SaVeR: Optimal Data Collection Strategy for Safe Policy Evaluation in Tabular MDP

Subhojyoti Mukherjee [1] **Josiah P. Hanna** [2] **Robert Nowak** [1]

Abstract

In this paper, we study safe data collection for the purpose of policy evaluation in tabular Markov decision processes (MDPs). In policy evaluation, we are given a *target* policy and asked to estimate the expected cumulative reward it will obtain. Policy evaluation requires data and we are interested in the question of what *behavior* policy should collect the data for the most accurate evaluation of the target policy. While prior work has considered behavior policy selection, in this paper, we additionally consider a safety constraint on the behavior policy. Namely, we assume there exists a known default policy that incurs a particular expected cost when run and we enforce that the cumulative cost of all behavior policies ran is better than a constant factor of the cost that would be incurred had we always run the default policy. We first show that there exists a class of intractable MDPs where no safe oracle algorithm with knowledge about problem parameters can efficiently collect data and satisfy the safety constraints. We then define the tractability condition for an MDP such that a safe oracle algorithm can efficiently collect data and using that we prove the first lower bound for this setting. We then introduce an algorithm SaVeR for this problem that approximates the safe oracle algorithm and bound the finite-sample mean squared error of the algorithm while ensuring it satisfies the safety constraint. Finally, we show in simulations that SaVeR produces low MSE policy evaluation while satisfying the safety constraint.

[1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, USA [2]Computer Sciences Department, University of Wisconsin-Madison, Madison, USA. Correspondence to: Subhojyoti Mukherjee <smukherjee27@wisc.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Reinforcement learning has emerged as a powerful tool for decision-making in a wide range of applications, from robotics (Ibarz et al., 2021; Agarwal et al., 2022) and game-playing (Szita, 2012) to autonomous driving (Kiran et al., 2021), web-marketing (Bottou et al., 2013), healthcare (Fischer, 2018; Yu et al., 2019) and finance (Hambly et al., 2021). However, in these applications, it is often necessary to first evaluate the decision-making policy before its long-term deployment in the real world. In fact, policy evaluation is a critical step in reinforcement learning, as it allows us to assess the quality of a learned policy and to check whether it can truly achieve the desired goal for the target task. One potential solution to this issue is off-policy evaluation (OPE) (Dudík et al., 2014; Li et al., 2015; Swaminathan et al., 2017; Wang et al., 2017; Su et al., 2020; Kallus et al., 2021; Cai et al., 2021). However, for OPE estimators there is no control over how the static dataset is generated, which could result in low accuracy estimates.

Hence, a natural idea is to actively gather the dataset using an adaptive behavior policy and thus increase accuracy in the evaluation of the target policy's value. In many real-world settings, the behavior policy itself must satisfy some side constraints (specific to the industry) (Wu et al., 2016) or safety constraints (Wan et al., 2022) while collecting the dataset. For instance, in web marketing, it is common to run an A/B test with safety constraints over a subset of all users before a potential new policy is deployed for all users (Kohavi and Longbotham, 2017; Tucker and Joachims, 2022). While testing autonomous vehicles it is quite natural to incorporate safety constraints in the behavior policy. So it is of great practical importance to ensure that our data collection rule is safe (Zhu and Kveton, 2022).

In this paper, we consider the question of optimal data collection for policy evaluation under safety constraints in the tabular reinforcement learning (RL) setting. Consider the following scenario that could arise in web marketing. Suppose we have a policy learned from offline data that has never been run in a real application. Moreover, we want this learned policy to be at least as good as a baseline policy that is already deployed in the application (Wu et al., 2016; Zhu and Kveton, 2021; 2022). Off-policy evaluation often

has high variance, so engineers may want to have some controlled deployment where the learned policy only makes decisions for some users before letting the policy make decisions for all users. We are motivated by how to make this controlled deployment as data-efficient and safe as possible. By *safe*, we mean that we want the expected return seen during data collection to remain close to the expected return under the baseline policy. A similar motivation can be found in Tucker and Joachims (2022). In this paper, we focus on finding a behavior policy that produces a minimal variance estimate while remaining safe. We can state this formally as follows: We are given a target policy, π, for which we want to estimate its value denoted by $V^\pi(s_1)$. To estimate $V^\pi(s_1)$ we will generate a set of K episodes where each episodic interaction ends after L timesteps. We denote the total available budget of samples as $n = KL$. Each episode is generated by following some behavior policy and collect the dataset $\mathcal{D}$. Let $Y_n^\pi(s_1)$ be the estimate of $V^\pi(s_1)$ computed from $\mathcal{D}$. Then our objective is to determine a sequence of behavior policies that minimizes error in the estimation of $V^\pi(s_1)$ defined as $\mathbb{E}_\mathcal{D}[(Y_n^\pi(s_1) - V^\pi(s_1))^2]$ subject to a *safety constraint* on the cost-value of the behavior policies (to be defined later) that must hold with high probability.

There is a growing body of literature studying this important problem of data collection for policy evaluation in both constrained and unconstrained setups. The work of Antos et al. (2008); Carpentier and Munos (2011; 2012); Carpentier et al. (2015); Fontaine et al. (2021); Mukherjee et al. (2022a; 2024) studies this problem in the bandit setting without any constraints under the finite sample regime. A common metric of performance that these works consider is the difference between the loss of the agnostic algorithm that does not know problem-dependent parameters, and the oracle loss (which has access to problem-dependent parameters). This metric is termed *regret* and these works show that in the bandit setting the regret of the agnostic algorithm scales as $\widetilde{O}(n^{-3/2})$ where $\widetilde{O}(\cdot)$ hides log factors. One might be tempted to just run the target policy π, build $\mathcal{D}$ and then estimate $Y_n^\pi(s_1)$. This is called *on-policy* data collection. However, these works show that the on-policy regret degrades at a much slower rate of $\widetilde{O}(n^{-1})$ compared to active agnostic algorithms. Hence, a natural question arises, can we achieve similar performance for policy evaluation in the MDP setup under a finite sample regime even when we must conform to safety constraints? Thus, the goal of our work is to answer the following questions:

> *1) Is there a class of MDPs where it is possible to incur a regret that degrades at a faster rate than $\widetilde{O}(n^{-1})$?* while satisfying safety constraints?
>
> *2) If the answer is yes to (1), can we design an adaptive algorithm (for this class of MDPs) to collect data for policy evaluation that does not violate the safety constraints (in expectation), and its regret degrades at a faster rate than $\widetilde{O}(n^{-1})$?*

In this paper, we answer these questions affirmatively. Regarding the first question, we state the tractability condition on the class of MDPs which enables the optimal behavior policy to gather data for policy evaluation without violating the safety constraint and suffer a regret of $\widetilde{O}(n^{-3/2})$. This condition leads to the first lower bound for this setting.

We also note that safe data collection for policy evaluation has also been studied in the bandit setting in Zhu and Kveton (2021; 2022). However, these works provide asymptotic guarantees whereas we are the first to provide finite-time regret guarantees when per-step constraints must be maintained in expectation. We also show that in the bandit setup, our method empirically outperforms the adaptive importance sampling based algorithms in these works. Our formulation is also related to constrained MDPs though we specify that the constraint must be satisfied *throughout learning* and not just by the final policy (Efroni et al., 2020; Vaswani et al., 2022). We discuss further related works in Appendix A.1.

Our main contributions are as follows:

(1) We formulate the problem of safe data collection for policy evaluation. We introduce the safety constraint such that at the end of n trajectories, the cumulative cost is above a constant factor of the baseline cost. To our knowledge, this is the first work to study this setting under such a safety constraint in the MDP setup with the goal of minimizing the estimate of the MSE of the target policy's expected reward.

(2) We then show that even in the special case of finite tree-structured MDPs the safe data collection for policy evaluation can be intractable. Then we come up with a condition on MDPs that enables any behavior policy to collect data without violating safety constraints. We also provide the first regret lower bound for the bandit and Tree MDP setting and show that it scales with $\Omega(n^{-3/2})$.

(3) We then consider an oracle strategy that knows the reward variances (problem-dependent parameter) of the reward distributions and derives its sampling strategy. We then introduce the agnostic algorithm **Sa**fe **V**arianc**e** **R**eduction (SaVeR) that does not know the problem-dependent parameters and show that its regret scales as $\widetilde{O}(n^{-3/2})$. We evaluate its performance against other baseline approaches and show that SaVeR reduces MSE faster while satisfying the safety constraint.

2. Preliminaries

We consider the standard finite-horizon Markov Decision process, $\mathcal{M}$, with both a reward and constraint function. Formally, $\mathcal{M}$, is a tuple $(\mathcal{S}, \mathcal{A}, P, R, C, \gamma, d_0, L)$, where $\mathcal{S}$

is a finite set of states, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a state transition function, R is the reward function (formalized below), C is the constraint function (formalized below), $\gamma \in [0, 1)$ is the discount factor, d_0 is the starting state distribution, and L is the maximum episode length. A (stationary) policy, $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, is a probability distribution over actions conditioned on a given state. We assume data can only be collected through episodic interaction: an agent begins in state $s_1 \sim d_0$ and then at each step t takes an action $a_t \sim \pi(\cdot|s_t)$ and proceeds to state $s_{t+1} \sim P(\cdot|s_t, a_t)$.

When the agent takes an action, a, in state, s, it receives both a reward $R \sim R(s, a)$ and a constraint value $C \sim C(s, a)$. We assume the transition model P is known but the reward distributions and constraint values are unknown. We define the reward value of a policy as: $V^\pi(s_1) := \mathbb{E}_\pi[\sum_{t=1}^n \gamma^{t-1} R_t]$, where $\mathbb{E}_\pi$ is the expectation w.r.t. trajectories sampled by following π from the initial state s_1. We define a constraint-value of π similarly: $V_c^\pi(s_1) := \mathbb{E}_\pi[\sum_{t=1}^n \gamma^{t-1} C_t]$. For simplicity, let the initial state distribution has probability mass on a single state s_1.

Our goal is to efficiently estimate $V^\pi(s_1)$ for a given policy π and this estimation requires data from the environment MDP. Past work has approached this problem by designing a sequence of *behavior policies* which are ran to produce informative data for evaluating π. However, in practical applications, it is often infeasible to simply run *any* behavior policy as doing so may violate domain constraints. We formalize this constraint by first assuming the existence of a safe *baseline policy*, π_0 that provides an acceptable constraint-value $V_c^{\pi_0}(s_1)$. Our objective is to determine a sequence of behavior policies, $\{\mathbf{b}_1, .., \mathbf{b}_K\}$, that will produce a set of K episodes that lead to the most accurate estimate of $V^\pi(s_1)$ subject to the constraint that the cumulative expected constraint-value $V_c^{\mathbf{b}}(s_1)$ always exceeds a fixed percentage of $V_c^{\pi_0}(s_1)$. We consider the objective:

$$\min_{\mathbf{b}} \mathbb{E}_{\mathcal{D}}[(Y_n^\pi(s_1) - V^\pi(s_1))^2] \tag{1}$$

$$\text{s.t.} \sum_{k'=1}^{k} V_c^{\mathbf{b}^{k'}}(s_1) \geq (1 - \alpha) k V_c^{\pi_0}(s_1) \text{ for all } k \in [K]$$

where $Y_n(s_1)$ is our estimate of $V^\pi(s_1)$, $\alpha \in (0, 1]$ is the risk parameter, and the expectation is over the collected data set $\mathcal{D}$. We also make the following simplifying assumption. We assume π_0 is deterministic, i.e., will only select one action in any given state. W.l.o.g., we give this action the index 0 and refer to it as the *safe action*. The entire action set is $\mathcal{A} = \{0, 1, \ldots, A\}$. This assumption is reasonable in applications where existing, safe policies were created through non-learning methods or manually designed.

For analysis, we will estimate $V^\pi(s_1)$ with a certainty-equivalence estimator. We define the random variable representing the estimated future reward from state s at time-step ℓ as $Y_n^\pi(s, \ell) := \sum_a \pi(a|s)\widehat{\mu}_n(s, a) + \gamma \sum_{s'} \widehat{P}_n(s'|s, a) Y_n^\pi(s', \ell + 1)$ where $Y_n^\pi(s, \ell + 1) := 0$ if $\ell \geq L$, and $\widehat{\mu}_n(s, a)$ is an estimate of $\mu(s, a)$, both computed from $\mathcal{D}$. Finally, the estimate of $V^\pi(s_1)$ is computed as $Y_n^\pi(s_1) := \sum_s d_0(s_1) Y_n^\pi(s_1, 0)$. Note that the total available budget of samples is n. We assume that there are K episodes and each episodic interaction terminates in at most L steps which implies $n = KL$.

We assume $V_c^{\mathbf{b}}(s_1)$ is known for $\mathbf{b} = \pi_0$ but not for any other policy. The constraint in (1) implies that the total constraint value over all deployed behavior policies should be above the total constraint value that can be obtained from the baseline policy π_0 till episode k with high probability. Observe that small values of α force the learner to be highly conservative, whereas larger α values correspond to a weaker constraint. A similar setting has been studied for policy improvement by Wu et al. (2016); Yang et al. (2021) for a variety of sequential decision-making settings. However, our objective is policy evaluation and we formulate a more general safety constraint in terms of $C(\cdot)$ while these prior works define the constraint in terms of $R(\cdot)$.

Similar to the recent works of Chowdhury et al. (2021); Ouhamma et al. (2023); Agarwal et al. (2019); Lattimore and Szepesvári (2020) we assume the reward function $R(s, a) = \mathcal{N}(\mu(s, a), \sigma^2(s, a))$, where $\mathcal{N}$ denotes a Gaussian distribution with mean $\mu(s, a)$ and variance $\sigma^2(s, a)$. Similarly we assume a constraint function $C(s, a) = \mathcal{N}(\mu^c(s, a), \sigma^{c,(2)}(s, a))$, where $\mu^c(s, a)$ and $\sigma^{c,(2)}(s, a)$ are the mean and variance of $\mathcal{N}(\cdot)$. Note that this sub-Gaussian distribution assumption is required only for theoretical analysis, whereas our algorithm works for any bounded reward and cost functions. We assume that we have bounded reward and constraint mean $\mu(s, a), \mu^c(s, a) \in [0, \eta]$ respectively. Finally, we define the MSE of a behavior policy $\mathbf{b}$ for the target policy π at the end of budget n as

$$\mathcal{L}_n(\pi, \mathbf{b}) = \mathbb{E}_{\mathcal{D}}[(Y_n^\pi(s_1) - V^\pi(s_1))^2] \tag{2}$$

where the expectation is over dataset $\mathcal{D}$ which is collected by $\mathbf{b}$. Our main objective is to minimize the cumulative regret $\mathcal{R}_n$ subject to the safety constraint defined in (1). To define $\mathcal{R}_n$ we first define the MSE of a safe oracle behavior policy $\mathbf{b}_*^k$ that collects the dataset $\mathcal{D}$ as $\mathcal{L}_n^*(\pi, \mathbf{b}_*^k)$. We will formally describe such oracle policies in Section 3. Then the regret $\mathcal{R}_n$ is defined as

$$\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*^k). \tag{3}$$

3. Intractability and Lower Bounds

In this section, we first define an oracle data collection strategy that ignores the constraints. We call this the unconstrained oracle. This oracle data collection algorithm can

reach a regret bound of $\widetilde{O}(n^{-3/2})$ in the unconstrained setting (Carpentier and Munos, 2012; Carpentier et al., 2015; Mukherjee et al., 2022a). We then show how data collection for policy evaluation under safety constraints in MDPs is challenging compared to standard policy improvement challenges in constrained MDPs (Efroni et al., 2020; Vaswani et al., 2022) as well as safe data collection for policy evaluation in bandits (Zhu and Kveton, 2021; Wan et al., 2022; Zhu and Kveton, 2022). To show this challenging aspect, we first discuss how the unconstrained oracle fails to satisfy the constraint and achieve the desired regret of $\widetilde{O}(n^{-3/2})$ in the constraint MDP setting. We then propose a safe variant of the oracle policy and finally, discuss a tractability condition that enables the safe oracle algorithm to achieve a regret bound of $\widetilde{O}(n^{-3/2})$.

3.1. Unconstrained Oracle

In this section, we discuss the unconstrained oracle data collection strategy that knows the variances of reward and constraint value but does not know the mean of either. Moreover, this oracle does not take into account the safety constraints in (1). After observing n samples (state-action-reward tuples), the oracle computes the estimate of $V^\pi(s_1^1)$ as $Y_n^\pi(s_1^1) = \sum_{a=1}^A \pi(a|s_1^1)\big(\widehat{\mu}_n(s_1^1,a) + \sum_{s_j^{\ell+1}} P(s_j^2|s_1^1,a)Y_n(s_j^2)\big)$. Note that we defined $Y_n^\pi(s,\ell)$ before, but now we use $Y_n^\pi(s)$ and assume the timestep is implicit in the state for this finite-horizon MDP. Mukherjee et al. (2022a) shows that in the unconstrained setting, to reduce the $\mathbf{Var}(Y_n^\pi(s_1^1))$ the optimal sampling proportion of the oracle for any state s_i^ℓ is:

$$\mathbf{b}_*(a|s_i^\ell) \propto \big(\pi^2(a|s_i^\ell)\big[\sigma^2(s_i^\ell,a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell,a)M^2(s_j^{\ell+1})\big]\big)^{\frac{1}{2}} \tag{4}$$

where, $M(s_j^\ell)$ is the normalization factor defined as follows:

$$M(s_i^\ell) = \sum_a \big(\pi^2(a|s_i^\ell)\big(\sigma^2(s_i^\ell,a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell,a)M^2(s_j^{\ell+1})\big)\big)^{\frac{1}{2}}. \tag{5}$$

Observe from the definition of $\mathbf{b}_*(a|s_i^\ell)$ that the optimal proportion in the terminal states, i.e. $\mathbf{b}_*(a|s_j^L)$, do not affect subsequent states and only depends on the target probability $\pi^2(a|s_i^\ell)$ and variance $\sigma^2(s_i^\ell,a)$. The key difference is in the non-terminal states, s_i^{L-1}, where the optimal action proportion, $\mathbf{b}_*(a|s_i^{L-1})$ depends on the expected terminal state normalization factor $M(s_j^L)$ where s_j^L is a state sampled from $P(\cdot|s_i^{L-1},a)$. The normalization factor, $M(s_j^L)$, captures the total contribution of state s_j^L to the variance of $Y_n^\pi(s_j^{L-1})$ and thus actions in the starting state must be chosen to 1) reduce variance in the immediate reward estimate and to 2) get to states that contribute more to the variance of the estimate. This observation is also noted in Mukherjee et al. (2022a). Finally, since $\mathbf{b}_*(a|s)$ also depends on $P(s'|s,a)$, it will put a low sampling proportion on actions a leading to such s' which has low transition probabilities.

3.2. Safe Oracle Algorithm for Safe Data Collection

The behavior policy defined in the previous section ignores the safety constraint and is thus inapplicable to our problem setting. In this section, we describe a safe variant of this oracle. We define a few notations before introducing the safe algorithm. Let $T_\ell^k(s,a) := \sum_{k'=1}^{k-1}\sum_{\ell'=1}^{\ell-1}\mathbf{1}\{S_{\ell'}^{k'}=s, A_{\ell'}^{k'}=a\}$ be the number of times (s,a) is visited before episode k. Let the mean reward estimate of (s,a) till episode k be computed as $\widehat{\mu}_\ell^k(s,a) := (T_\ell^k(s,a))^{-1}\sum_{k'=1}^{k-1}\sum_{\ell'=1}^{\ell-1}\mathbf{1}\{S_{\ell'}^{k'}=s, A_{\ell'}^{k'}=a\}R_{\ell'}^{k'}$, where $R_{\ell'}^{k'}$ is the observed reward. Similarly define the constraint-values estimate $\widehat{\mu}_{c,\ell}^k(s,a)$ based on constraint value C_ℓ^k. Define the confidence interval at the timestep L of k-th episode as $\beta_L^k(s,a) := L\sqrt{\log(SAn(n+1))/T_L^k(s,a)}$ (Agarwal et al., 2019).

Let $Y_{c,L}^{\mathbf{b}^k}(s_1^1) = \sum_{a=1}^A \mathbf{b}^k(a|s_1^1)\big(\widehat{\mu}_{c,L}^k(s_1^1,a) + \sum_{s_j^{\ell+1}} P(s_j^2|s_1^1,a)Y_{c,L}^{\mathbf{b}^k}(s_j^2)\big)$ denote the empirical estimate of $V_c^{\mathbf{b}^k}(s_1^1)$ at the end of the k-th episode, and $\widehat{\mu}_{c,L}^k(s,a)$ is the empirical estimate of $\mu^c(s,a)$ at the end of the k-th episode. Note that the oracle algorithm knows the variances of reward $R(\cdot)$ and constraint-value $C(\cdot)$. Using this knowledge, it maintains a safety budget $\widehat{Z}_L^{k-1}$ where $\widehat{Z}_L^{k-1} := \sum_{k'=1}^{k-1}(Y_{c,L}^{\mathbf{b}^{k'}}(s_1^1) - \beta_L^{k'}(s,a)) - (1-\alpha)(k-1)V_c^{\pi_0}(s_1^1)$ is the safety budget at the end the $k-1$-th episode. The $\underline{Y}_{c,L}^{\mathbf{b}^k}(s_1^1) = Y_{c,L}^{\mathbf{b}^k}(s_1^1) - \beta_L^k(s,a)$ is the lower confidence bound to the $Y_{c,L}^{\mathbf{b}^k}(s_1^1)$.

Exploration policy π_x: We require an exploration policy π_x as the oracle algorithm needs a good estimation of the constraint-value $\mu^c(s,a)$ and following the oracle proportion $\mathbf{b}_*(a|s)$ may not lead to a good estimation of $\mu^c(s,a)$. This exploration policy should ensure with high probability that the estimation error of $\mu^c(s,a)$ is low in each (s,a) for which $\pi(a|s) > 0$ and can be an optimal design based policy like PEDEL that explores the state space informatively (Wagenmaker and Jamieson, 2022) or other exploration policies (e.g., Dann et al. (2019); Ménard et al. (2020); Uehara et al. (2021)).

We now state the following safe oracle algorithm: At the k-th episode run the policy

$$\mathbf{b}_*^k = \begin{cases} \mathbf{b}_*, & \text{if } \widehat{Z}_L^{k-1} \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}_L^{k-1} < 0 \\ \pi_x, & \text{if } \widehat{Z}_L^{k-1} \geq 0, k \leq \sqrt{K} \end{cases}. \tag{6}$$

The safe oracle algorithm in (6) alternates between the op-

timal oracle policy $\mathbf{b}_*$ in (4) when the safety budget $\widehat{Z}_L^{k-1}$ at the start of the episode k is greater than 0, otherwise it falls back to running the baseline policy π_0. Additionally, the safe oracle conducts forced exploration for at most $\sqrt{K}$ episodes when $\widehat{Z}_L^{k-1} \geq 0$ using the exploration policy π_x to estimate $\mu^c(s,a)$. This is because following the oracle proportion $\mathbf{b}_*$ in (4) that samples high variance state-action tuples may not lead to a good estimate of $\mu^c(s,a)$.

3.3. An Intractable MDP

In this section, we now show that there exist MDPs where even a safe oracle algorithm may not be able to reach the desired $\widetilde{O}(n^{-3/2})$ regret bound. We then introduce the tractability condition which depends on the budget as the $\mathbf{b}_*$ needs to be run sufficient number of times to reach a regret of $\widetilde{O}(n^{-3/2})$. So a more benign MDP allows one to run $\mathbf{b}_*$ most of the time whereas a less benign MDP allows you to play $\mathbf{b}_*$ less. Hence tractability depends on the budget being sufficiently large and also depends on properties of the MDP and the risk parameter α. To show this challenging aspect of safe data collection, we first define a Tree MDP. Using Tree MDPs to analyze the hardness of learning in MDPs and deriving lower bounds is common in the literature (Jiang and Li, 2016; Weisz et al., 2021; Wagenmaker et al., 2022; Jin et al., 2022). The tree MDP is defined as follows:

Definition 3.1. (Tree MDP) An MDP is a discrete tree MDP $\mathcal{T} \subset \mathcal{M}$ in which: **(1)** There are L levels indexed by ℓ where $\ell = 1, 2, \ldots, L$. **(2)** Every state is represented as s_i^ℓ where ℓ is the level of the state s indexed by i. **(3)** The transition probabilities are such that one can only transition from a state in level ℓ to one in level $\ell+1$ and each non-initial state can only be reached through one other state and only one action in that state. Formally, $\forall s', P(s'|s,a) \neq 0$ for only one state-action pair s, a and if s' is in level $\ell+1$ then s is in level ℓ. Finally, $P(s_j^{L+1}|s_i^L, a) = 0, \forall a$. **(4)** For simplicity, we assume that there is a single starting state s_1^1 (called the root). It is easy to extend our results to multiple starting states with a starting state distribution, d_0, by assuming that there is only one action available in the root that leads to each possible start state, s, with probability $d_0(s)$. The leaf states are denoted as s_i^L. **(5)** The interaction stops after L steps in state s_i^L after taking an action a.

Proposition 1. *Fix an arbitrary $n > 0$. Then there exists an environment where no algorithm (including the safe oracle $\mathbf{b}_*^k$) can be run that will result in a regret $\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}_k^*) - \mathcal{L}_n^*(\pi, \mathbf{b}_*)$ of $\widetilde{O}(n^{-3/2})$ while satisfying the safety constraint, where $\mathbf{b}_*$ is the unconstrained oracle.*

Proof (Overview) We first construct a worst-case 3 armed bandit environment (MDP with single state) such that $\mu^c(0) = 0.5$, $\mu^c(1) = 0.5 + \alpha$, $\mu^c(2) = 0$ and variance of $\sigma^{r,(2)}(0) = 0.001$, $\sigma^{r,(2)}(1) = 0.001$ and $\sigma^{r,(2)}(2) = 0.25$. So action $\{2\}$ has low constraint value (unsafe) but has high variance. So the safe oracle policy must sample the action 2 a large number of times to reach a regret of $\widetilde{O}(n^{-3/2})$. However, since action $\{2\}$ is unsafe, the safe oracle has to sample baseline action 0 a sufficient number of times to accrue some safety budget. Combining these two observations we show that achieving a regret rate of $\widetilde{O}(n^{-3/2})$ is impossible. The full proof is in Appendix B.

The key reason the above environment is intractable is that some trajectories taken by safe oracle has very less constraint value associated with them, compared to the trajectory taken by the baseline policy. To rule out such pathological MDPs, we define the *tractability* condition as follows: Let $\mathbf{b}^-$ be any behavior policy that minimizes $V_{\mathbf{b}}^c(s_1)$. Define $V_{\mathbf{b}^-}^c(s_1)$ as the value of the policy $\mathbf{b}^-$ starting from state s_1. This policy $\mathbf{b}^-$ suffers a value $V_{\mathbf{b}^-}^c(s_1)$ that is lower than any other behavior policy $\mathbf{b}$. So this policy $\mathbf{b}^-$ can be thought of as the worst possible behavior policy that can be followed by the agent during an episode. Then the tractability condition states that

$$\sqrt{n} \geq \frac{\frac{1}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right)}{\frac{C_\sigma}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right) - 1} \tag{7}$$

where $C_\sigma \in (0,1)$ is a MDP dependent parameter that depends on the reward variance of state-action pairs such that $\frac{C_\sigma}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right) - 1 > 0$. The quantity $C_\sigma = \max_{s,a} \frac{\mathbf{b}_*(a|s)}{M(s)}$ where $\mathbf{b}_*(a|s)$ and $M(s)$ are defined in (4) and (5) respectively. So $C_\sigma \in (0,1)$ and it captures the worst case trajectory that can be followed by $\mathbf{b}_*$.

This condition gives us (1) the lower bound to the budget n to run the behavior policy $\mathbf{b}^-$ to achieve a regret bound of $\widetilde{O}(n^{-3/2})$ and satisfy the safety constraint; (2) $V_{\mathbf{b}^-}^c(s_1) < V_{\pi_0}^c(s_1)$ so that the RHS is positive, (3) depends on the reward variance of state action pairs in the MDP so that $\frac{C_\sigma}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right) - 1 > 0$, and (4) for smaller α (high risk) the R.H.S increases which increases the required budget n. We further discuss how this condition in (7) is derived in Remark B.1. Then we define the following assumption.

Assumption 3.2. (Tractability) We assume a sufficiently large budget n and an MDP $\mathcal{M}$ that satisfies the constraint in (7). We call such an MDP $\mathcal{M}$ *tractable*.

Assumption 3.2 ensures that even the worst possible behavior policy $\mathbf{b}^-$ that can reach a regret of $\widetilde{O}(n^{-3/2})$ has sufficient budget n to satisfy the safety constraint. Moving forward, we will define regret relative to this safe oracle $\mathbf{b}_*^K$ instead of the unconstrained oracle. Furthermore, we assume tractability in (3.2) such that the safe oracle decreases MSE at a comparable rate to the unconstrained oracle $\mathbf{b}_*$. Define the reward regret as $\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*^k)$

where $\mathcal{L}_n^*(\pi, \mathbf{b}_*^k)$ is the safe oracle MSE, and $\mathcal{L}_n(\pi, \mathbf{b})$ is the agnostic algorithm MSE that does not know reward or constraint-value variances. Now we present the first general lower bound theorem for the safe data collection strategy in MDPs.

Theorem 1. ***(Lower Bounds)*** *Let $\pi(a|s) = \frac{1}{A}$ for each state $s \in \mathcal{S}$. Under Assumption 3.2 the regret $\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*^k)$ is lower bounded by*

$$\mathbb{E}\left[\mathcal{R}_n\right] \geq \begin{cases} \Omega\left(\max\left\{\frac{A^{1/3}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 A^{2/3}}{n^{3/2}}\right)\right\}\right), \textbf{(MAB)} \\ \Omega\left(\max\left\{\frac{\sqrt{SAL^2}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 SAL^2}{n^{3/2}}\right)\right\}\right) \textbf{(MDP)} \end{cases}$$

where, $\Delta_0 = V_c^{\mathbf{b}_*}(s_1^1) - V_c^{\pi_0}(s_1^1)$ and $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)$ is the hardness parameter.

Discussion: Theorem 1 shows that in the constrained setting the lower bound scales as $\Omega(H_{*,(1)}^2 n^{-3/2})$. Note that we can recover the lower bound for the unconstrained setting using this result. In the unconstrained bandit setting the bound scales as $O\left(A^{1/3} n^{-3/2}\right)$ which matches the lower bound of Carpentier and Munos (2012) (see their Theorem 5). We also establish the first lower bound for the unconstrained setting in data collection for policy evaluation in the tabular MDP setup that scales as $O\left(\sqrt{SAL^2} n^{-3/2}\right)$. The $H_{*,(1)}$ captures the hardness in learning in the MDP and consists of the gap Δ_0, $V_c^{\pi_0}(s_1^1)$ and α. Note that $H_{*,(1)}$ increases with α, and the Δ_0 captures how much constraint value the $\mathbf{b}_*$ can obtain compared to π_0. Finally, the smaller value of π_0 increases the hardness as the π_0 has to be run more times so that the safety constraint is not violated.

Proof (Overview) We first build two deterministic tree MDPs $\mathcal{T}$ and $\mathcal{T}'$ which differ in the variances at only one state. This leads to different optimal oracle behavior policies in $\mathcal{T}$ and $\mathcal{T}'$. Then using the divergence decomposition lemma for MDPs from Garivier and Kaufmann (2016); Wagenmaker et al. (2022) we show in Lemma C.6 that in $\mathcal{T}$ the regret lower bound scales as $\Omega(\sqrt{SAL^2 \log(n)}/n^{3/2})$. Next, we follow a reduction-based proof technique to prove the reward regret lower bound in the constrained setting. Consider any sequential decision-making problem $\mathfrak{A}$ (for instance a multi-armed bandit problem, tabular RL) such that there exists a problem-dependent constant $\xi \in \mathbb{R}$ that only depends on on the number of actions in bandits, or state-action-horizon in tabular RL. Then for a large budget n and any algorithm we have from Lemma C.5 and Lemma C.6 that $\mathbb{E}[\mathcal{R}_n] \geq \frac{\xi}{n^{3/2}}$ for an MDP dependent parameter ξ. Then we lower bound how many times under the budget n the algorithm can run the baseline policy. This is lower bounded in step 2 as $\mathbb{E}[\mathcal{R}_n] \gtrsim \min\left\{\frac{\xi}{n^{3/2}}, \frac{(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)^2 \xi^2}{(\alpha V_c^{\pi_0}(s_1^1))^2 n^{3/2}}\right\}$. We finish off the proof by noting that the quantity $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)$ is the hardness parameter when $\pi(a|s) = 1/A$, and substituting the value of $\xi = A^{1/3}$ for bandits (Lemma C.5) and $\xi = \sqrt{SAL^2}$ for $\mathcal{T}$ (Lemma C.6). Since $\mathcal{T} \subset \mathcal{M}$, this result is a lower bound to $\mathcal{M}$ as well. The full proof is in Appendix C. ■

4. Agnostic Algorithm for Safe Policy Evaluation

In this section, we introduce the more realistic agnostic algorithm that does not know the mean and variances of the reward and constraint values of the actions. We then analyze this algorithm and establish its finite-time MSE. We call this algorithm **Sa**fe **V**arianc**e** **R**eduction algorithm (abbreviated as SaVeR) as it reduces the variance of the estimated value of the target policy by following (4) while simultaneously satisfying the safety constraint (1) with high probability.

We introduce a few notations before presenting the algorithm. Define the upper confidence bound on the empirical reward variance as $\overline{\widehat{\sigma}}_L^k(s,a) := \widehat{\sigma}_L^k(s,a) + \beta_L^k(s,a)$, where $\beta_L^k(s,a)$ is the confidence interval defined in Section 3.1. We define the empirical sampling proportion for an arbitrary state-action (s_i^ℓ, a) as $\widehat{\mathbf{b}}_\ell^k(a|s_i^\ell)$. Define the policy $\widehat{\mathbf{b}}_{*,\ell}^k(a|s_i^\ell)$ as similar to $\mathbf{b}_*(a|s_i^\ell)$ defined in (4), but it uses plug-in estimate $\overline{\widehat{\sigma}}_\ell^k(s,a)$ instead of $\sigma_\ell^k(s,a)$. This is because the agnostic algorithm does not know the reward and constraint-value variances. We define $\widehat{Z}_L^{k-1}$ similar to (6). Finally, we define our algorithm, SaVeR, as follows: At episode k run the policy:

$$\widehat{\mathbf{b}}^k = \begin{cases} \widehat{\mathbf{b}}_*^k & \text{if } \widehat{Z}^{k-1} \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}^{k-1} < 0 \\ \pi_x & \text{if } \widehat{Z}^{k-1} \geq 0, k \leq \sqrt{K} \end{cases} \tag{8}$$

where $\widehat{\mathbf{b}}_*^k$ for the episode k is defined as follows: For each timestep $\ell = 1, 2, \ldots, L$ sample action $A_\ell^k = \arg\max_a \frac{\widehat{\mathbf{b}}_*^k(a|s_j^\ell)}{T_\ell^k(s_j^\ell, a)}$, where $\widehat{\mathbf{b}}_*^k(a|s_j^\ell)$ is the plug-in estimate of $\mathbf{b}_*(a|s_j^\ell)$ as defined in (4). SaVeR alternates between the exploration policy π_x, plugin optimal policy $\widehat{\mathbf{b}}_*^k$, and baseline policy based on the safety budget $\widehat{Z}^k$ and the number of episodes K. In contrast to (8) the oracle policy in (6) uses the true oracle proportions $\mathbf{b}^*$ when $\widehat{Z}^{k-1} \geq 0, k > \sqrt{K}$. Also, observe that the action selection rule ensures that the ratio $\widehat{\mathbf{b}}_{*,\ell}^k(a|s)/T_\ell^k(s,a) \approx 1$. It is a deterministic action selection rule and thus avoids inadvertently violating the safety constraint due to random sampling from the optimal proportions $\widehat{\mathbf{b}}_\ell^k(a)$. Now we formally state the SaVeR for the tree MDP. At every episode $k \in [K]$ it generates a sampling history $\mathcal{H}^k := \{S_\ell^k, A_\ell^k, R(S_\ell^k, A_\ell^k), C(S_\ell^k, A_\ell^k)\}_{\ell=1}^L$ by selecting A_ℓ^k according to (8) and appends it to the dataset $\mathcal{D}$. After observing the feedback it updates the model param-

eters and estimates $\widehat{\mathbf{b}}_1^{k+1}(a|s)$ for each s, a. It returns the dataset $\mathcal{D}$ to evaluate π. The pseudocode is in Algorithm 1.

Algorithm 1 **Sa**fe **V**ariance **R**eduction (SaVeR) for $\mathcal{T}$

Input: Risk Parameter $\alpha > 0$, target policy π.
Output: Dataset $\mathcal{D}$.
Initialize $\mathcal{D} = \emptyset$, $\widehat{\mathbf{b}}_1(a|s)$ uniform over all actions.
for $k = 1, 2, \ldots, K$ **do**
  for $\ell = 1, 2, \ldots, L$ **do**
    Get $\mathcal{H}^k := \{S_\ell^k, A_\ell^k, R(S_\ell^k, A_\ell^k), C(S_\ell^k, A_\ell^k)\}_{\ell=1}^L$ by selecting $\mathbf{b}^k$ according to (8).
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathcal{H}^k, \widehat{\mathbf{b}}^k)\}$
    Update model parameters and estimate $\widehat{\mathbf{b}}_1^{k+1}(a|s)$ for each s, a
  end for
end for
Return Dataset $\mathcal{D}$ to evaluate policy π.

We now present a theorem that gives the MSE of the agnostic algorithm SaVeR in the tree MDP in the following theorem. We define the problem complexity parameters $M = \sum_{\ell=1}^L \sum_{s_j^\ell} M(s_j^\ell)$ summed over all stated $s \in [S]$. Define predicted agnostic constraint violation

$$\mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) := \sum_{k=1}^K \mathbb{I}\{\widehat{Z}^k < 0\}$$

when taking actions according to (8). For scalars $x, y \in \mathbb{R}$ define $\min^+(x, y) := |\min(x, y)|$. Define the problem complexity parameter $H_{*,(2)} = \sum_{\ell=1}^L \sum_{s_j^\ell} H_{*,(2)}(s_j^\ell)$ where

$$H_{*,(2)}(s_j^\ell) = \frac{1}{\alpha \mu^c(s_j^\ell, 0)} \sum_{a \in \mathcal{A} \setminus \{0\}} \pi(a|s_j^\ell)\sigma(s_j^\ell, a) \min^+ \{\Delta_c(s_j^\ell, a), \Delta_c(s_j^\ell, 0) - \Delta_c(s_j^\ell, a)\}\}. \quad (9)$$

Remark 4.1. The quantity $H_{*,(2)}(s_j^\ell)$ signifies the total cost of maintaining the safety constraint at state s_j^ℓ by sampling action 0 instead of sampling based on $\pi(a)\sigma(a)$. Observe that $\Delta_c(s_j^\ell, 0) - \Delta_c(s_j^\ell, a) = \mu_c(s_j^\ell, a) - \mu_c(s_j^\ell, 0)$. So $\min^+\{\Delta_c(s_j^\ell, a), \Delta_c(s_j^\ell, 0) - \Delta_c(s_j^\ell, a)\}$ depends on how close is the action a to the best cost action $\mu^{*,c}(s_j^\ell)$ or the baseline action 0. Also observe that because of the $\min^+$ operator, this quantity cannot be 0. Further, observe that the gap is weighted by $\pi(a|s_j^\ell)\sigma(s_j^\ell, a)$ signifying that actions with low variance and target probability contribute less to the constraint violation MSE. Also, observe that higher risk setting ($\alpha \to 0$) leads to higher $H_{*,(2)}(s_j^\ell)$. Finally, it can be easily verified that $H_{*,(2)} > H_{*,(1)}$.

Now we present a theorem that we will use to bound the regret of SaVeR in Tree MDP $\mathcal{T}$ under Assumption 3.2.

Theorem 2. ***(informal)*** *The MSE of the SaVeR in $\mathcal{T}$ for $\frac{n}{\log(SAn(n+1)/\delta)} \geq O((LSA^2)^2 + \frac{SA}{\Delta_{\min}^{c,(2)}} + \frac{1}{4H_{*,(2)}^2})$ is bounded by $\mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k) \leq \widetilde{O}\big(\frac{M^2(s_1^1)}{n} + \frac{M^2(s_1^1)}{n}(MLSA^2 + H_{*,(2)})^2 + \frac{(LSA^2)^2 H_{*,(2)}^2 M^2}{\min_s \mathbf{b}^{*,k,(3/2)}(s) n^{3/2}}\big)$ with probability $(1-\delta)$. The total predicted constraint violations are bounded by $\mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) \leq \widetilde{O}\big(\frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + LSA^2 + \frac{(LSA^2)^2 H_{*,(2)}^2 M^2}{n^{1/2}}\big)$ with probability $(1-\delta)$, where $M_{\min} := \min_s M(s)$.*

Discussion: In Theorem 2 the first quantity upper bounding $\mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k)$ is denoted as the *safe MSE* when the safety budget $\widehat{Z}^k \geq 0$ and scales as $M^2(s_1^1)/n$. The second quantity is denoted as the *unsafe MSE* which is accumulated due to constraint violation ($\widehat{Z}^k < 0$) and sampling of the safe action 0. Finally, the third quantity is the MSE suffered due to estimation error of the variances $\sigma^2(s, a)$. Comparing the result of the Theorem 2 with the unconstrained setting of Mukherjee et al. (2022a) we have the additional quantity of $(MLSA^2 + H_{*,(2)})^2/n$ where $H_{*,(2)}$ is the problem-dependent quantity summed over all states. Observe that if all actions are safe then we have that $\mathcal{L}_n^*(\pi, \widehat{\mathbf{b}}^k) = M^2(s_1^1)/n$ which recovers the MSE of the unconstraint setting in Carpentier and Munos (2011; 2012); Carpentier et al. (2015); Mukherjee et al. (2022a).

Proof (Overview) The agnostic SaVeR does not know the reward variances. The sampling rule in (8) ensures that the good variance event $\xi_{v,K}$ defined in (18) (step 2) holds such that SaVeR has good estimates of reward variances. Then, note that in the tree MDP $\mathcal{T}$ we have a closed form expression of $\mathbf{b}_*(s_j^\ell|a)$. We divide the total budget $n = n_f + n_u$ where n_f are the samples allocated when safety budget $\widehat{Z}^k \geq 0$. The n_f samples are also used by the exploration policy π_x to ensure a good estimate of the constraint means as stated in the event $\xi_{c,K}$ (17). This is ensured by π_x and noting that $n > SA\log(1/\delta)/\Delta_{c,\min}^2$. The remaining samples from n_f are allocated for reducing the MSE by sampling according to $\arg\max_a(\mathbf{b}_*(a|s)/T_\ell^k(s, a))$. We again prove an upper and lower bound to $T_n(s, a)$ in (27) in step 4 and (28) in step 5. Finally using Lemma A.1 we can bound the MSE for the duration n_f for all actions $a \in \mathcal{A} \setminus \{0\}$ for each state s_j^ℓ in step 6. Now for an upper bound to constraint violations, we use the gap $\Delta_c^\alpha(s, a) := (1-\alpha)\mu_{c,0}(s, a) - \mu_c(s, a)$ to bound how much each $a \in \mathcal{A} \setminus \{0\}$ in s_j^ℓ is underpulled and their pulls replaced by action $\{0\}$ weighted by $\pi(a|s_j^\ell)\sigma(s_j^\ell, a)$. This is captured by $H_{*,(2)}(s)$. Summing over all s, and horizon L gives the upper bound to the violations as shown in step 7. Finally, we also show a lower bound to constraint violations to bound the MSE for the duration when actions $a \in \mathcal{A} \setminus \{0\}$ are underpulled. This is shown in steps 8 and 9 where we equate the safety budget to 0 to obtain a lower bound to $T_n(s_j^\ell, 0)$ for each state s_j^ℓ. Combining everything in step 10 gives the result. The proof is in Appendix D. ■

Note that we do not have a closed-form solution to $\mathbf{b}_*^k$ that both minimizes MSE as well as upholds (1) for all $k \in [K]$ (as opposed to Carpentier and Munos (2011); Mukherjee et al. (2022b)). Therefore, we now define two additional notions of regret. The first is the regret defined as $\overline{\mathcal{R}}_n = \mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{L}}_n^*(\pi, \mathbf{b}_*^k)$ where $\overline{\mathcal{L}}_n^*(\pi, \mathbf{b}_*^k)$ is the upper bound to the safe oracle MSE. The second is the constraint regret defined as follows: $\overline{\mathcal{R}}_n^c = \mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{C}}_n^*(\pi, \mathbf{b}_*^k)$ where $\overline{\mathcal{C}}_n^*(\pi, \mathbf{b}_*^k)$ is the upper bound to the oracle constraint violations. Note that the oracle knows the variances of reward and constraint-values for all state-action tuples (but does not know the mean of either). The following corollary bounds SaVeR regret.

Corollary 1. *Under Assumption 3.2, the constraint regret of SaVeR is bounded by* $\overline{\mathcal{R}}_n^c \leq O\big(\frac{\log(n)}{n^{1/2}}\big)$ *and the regret is bounded by* $\overline{\mathcal{R}}_n \leq O\big(\frac{\log(n)}{n^{3/2}}\big)$.

The proof is in Appendix E.1 and directly follows from Theorem 2, and Proposition 2. In Proposition 2 in Appendix E we prove the MSE upper bound of the oracle. Observe, that the regret decreases at a rate of $\widetilde{O}(n^{-3/2})$, faster than the rate of decrease of on-policy MSE of $\widetilde{O}(n^{-1})$. Thus, we have been able to answer the second main question of this paper affirmatively. We also state a constraint and regret upper bound in the bandit setting in Corollary 2 in Appendix E.1. Also, observe that our upper bound matches the rate in the lower bound shown in Theorem 1.

5. Extension to DAG

In this section, we approximate the solution in $\mathcal{T}$ to DAG $\mathcal{G}$ and formulate the safe algorithm for policy evaluation. We first define the DAG MDP in the following definition.

Definition 5.1. (DAG MDP) A DAG MDP follows the same definition as the tree MDP in Definition 3.1 except $P(s'|s,a)$ can be non-zero for any s in layer ℓ, s' in layer $\ell+1$, and any a, i.e., one can now reach s' through multiple previous state-action pairs.

Then we state the following lemma from Mukherjee et al. (2022a).

Lemma 5.2. ***(Proposition 3 of Mukherjee et al. (2022a))*** *Let $\mathcal{G}$ be a 3-depth, A-action DAG defined in Definition 5.1. The minimal-MSE sampling proportions $\mathbf{b}_*(a|s_1^1), \mathbf{b}_*(a|s_j^2)$ depend on themselves such that $\mathbf{b}(a|s_1^1) \propto f(1/\mathbf{b}(a|s_1^1))$ and $\mathbf{b}(a|s_j^2) \propto f(1/\mathbf{b}(a|s_j^2))$ where $f(\cdot)$ is a function that hides other dependencies on variances of s and its children.*

The Lemma 5.2 (Mukherjee et al., 2022a) shows that one cannot derive a closed-form solution to $\mathbf{b}_*$ in $\mathcal{G}$ because of the existence of multiple paths to the same state resulting in a cyclical dependency. Note that in $\mathcal{T}$ there is only a single path to each state and this cyclical dependency does not arise. If we ignore the multiple path problem, we can approximate the optimal sampling proportion in $\mathcal{G}$ by using the tree formulation in the following way: At every time t during an episode k call the Algorithm 2 to estimate $M_0(s)$ where $M_{t'}(s) \in \mathbb{R}^{L\times|\mathcal{S}|}$ stores the expected standard deviation of the state s at iteration t'. After L such iteration we use the value $B_0(s)$ to estimate $\mathbf{b}(a|s)$ as follows:

$$\mathbf{b}_*(a|s) \propto \sqrt{\pi^2(a|s)\Big[\sigma^2(s,a) + \gamma^2 \sum_{s'} P(s'|s,a) M_0^2(s)\Big]}.$$

Note that for a terminal state s we have the transition probability $P(s'|s,a) = 0$ and then the $b(a|s) = \pi(a|s)\sigma(s,a)$. This iterative procedure follows from the tree formulation in Lemma A.2 and is necessary in $\mathcal{G}$ to take into account the multiple paths to a particular state. Algorithm 2 gives pseudocode for this procedure which takes inspiration from value-iteration for the episodic setting.

Algorithm 2 Estimate $B_0(s)$ for $\mathcal{G}$

Initialize $B_L(s) = 0$ for all $s \in \mathcal{S}$
for $t' \in L-1, \ldots, 0$ **do**
  $B_{t'}(s) = \sum_a \big(\pi^2(a|s)\big(\sigma^2(s,a) + \gamma^2 \sum_{s'} P(s'|s,a) B_{t'+1}^2(s)\big)\big)^{\frac{1}{2}}$
end for
Return B_0.

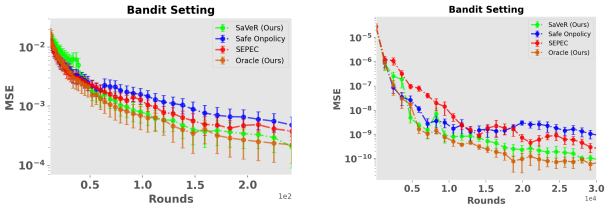
Finally, the safe algorithm in $\mathcal{G}$ can be stated as follows: At episode k

$$\text{Play } \mathbf{b}^k = \begin{cases} \pi_e & \text{if } \widehat{Z}^k \geq 0, k \leq \sqrt{K} \\ \pi_{\widehat{\mathbf{b}}^k} & \text{if } \widehat{Z}^k \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}^k < 0 \end{cases} \tag{10}$$

where $\pi_{\widehat{\mathbf{b}}^k}$ for the episode k is defined as follows: For each time $\ell = 1, 2, \ldots, L$ sample action $A_\ell^k = \arg\max_a \frac{\widehat{\mathbf{b}}^k(a|s_j^\ell)}{T_\ell^k(s_j^\ell, a)}$, where $\widehat{\mathbf{b}}^k(a|s_j^\ell)$ is the plug-in estimate of $\mathbf{b}_*(a|s_j^\ell)$ that is obtained using Algorithm 2.

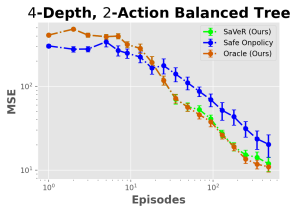
6. Experiments

In this section, we show numerical experiments validating our theoretical results. The full experimental details and numerical results are in Appendix G. We test the oracle, and SaVeR algorithm and introduce a method that we call safe on-policy. The safe on-policy algorithm follows the target policy π when the safety budget is positive and plays baseline policy π_0 when the safety budget is negative. We also test against the SEPEC (Wan et al., 2022) algorithm for the bandit setting which uses importance sampling to safely collect data for policy evaluation. Note that the bandit setting consists of a single state and every episode K consists of a

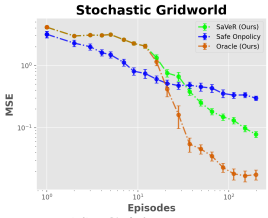


(a) Bandit setting

(b) Movielens setting

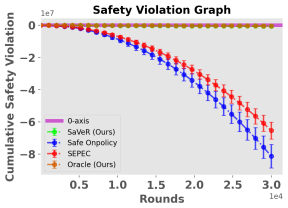


(c) Tree MDP

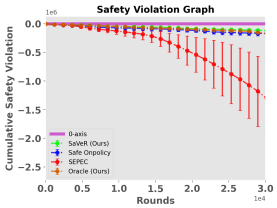


(d) Grid MDP

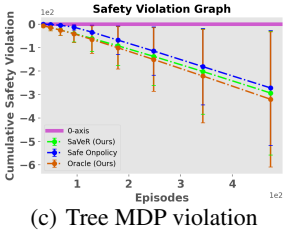
Figure 1. MSE in different settings. The vertical axis (log-scaled) gives MSE and the horizontal axis is the number of episodes (or rounds for bandits). Confidence bars show one standard error.



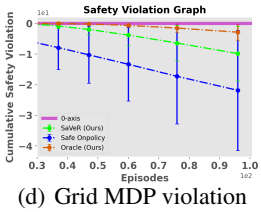
(a) Bandit violation



(b) Movielens violation



(c) Tree MDP violation



(d) Grid MDP violation

Figure 2. The vertical axis gives cumulative constraint violation and the horizontal axis is the number of episodes/rounds. The 0-axis is shown in pink. A safe algorithm has its plot below the 0-axis with the plot showing the cumulative unsafe budget.

single timestep $L = 1$. Figure 1 shows the MSE obtained by each algorithm for a varying number of episodes. In Figure 2, we show that all algorithms respect the constraint but that the oracle and SaVeR are not excessively conservative.

Experiment 1 (Bandit): We implement a general bandit environment with $A = 11$ and show that SaVeR achieves lower MSE than SEPEC and safe on-policy algorithm as the number of rounds increases. The performance is shown in Figure 1(a). From Figure 2(a) we see that SaVeR, and oracle do not oversample the safe action but allocate the right amount to be just safe. They allocate more samples to reduce the MSE, whereas the safe on-policy and SEPEC over-sample the safe action instead of focusing on reducing the MSE.

Experiment 2 (Movielens): We conduct this experiment on the real-life Movielens 1M dataset (Lam and Herlocker, 2016) for $A = 30$ actions and show that SaVeR achieves lower MSE than safe on-policy and SEPEC algorithm as the number of rounds increases. The performance is shown in Figure 1(b). From Figure 2(b), we see that SaVeR and oracle SaVeR, and the oracle do not oversample the safe action compared to SEPEC.

Experiment 3 (Tree): We experiment with a 4-depth 2-action deterministic tree MDP consisting of 15 states. With increasing episodes SaVeR reaches lower MSE than safe on-policy and eventually matches the oracle's MSE in Figure 1(c). In Figure 2(c) the SaVeR and oracle run the baseline policy almost similar number of times compared to the safe on-policy.

Experiment 4 (Gridworld): This setting consist of a 4×4 stochastic gridworld of 16 grid cells. We point out that Gridworld has a DAG structure (due to the finite horizon) which violates the tree structure assumption under which the oracle and SaVeR bounds were derived. Nevertheless, both SaVeR and oracle reach lower MSE with increasing episodes compared to safe onpolicy in Figure 1(d). We use (10) to estimate $\widehat{\mathbf{b}}$ in this setting. In Figure 2(d) we see that SaVeR allocates more samples to reduce the MSE, whereas the safe on-policy runs the baseline policy more instead of focusing on reducing the MSE.

7. Conclusions

In this paper, we studied the question of how to take action to build a dataset for minimal-variance policy evaluation of a fixed target policy under a safety constraint (1). We developed a theoretical foundation for data collection in policy evaluation by showing that there exists a class of MDPs (namely tree-structured MDPs $\mathcal{T}$) where safe policy evaluation is intractable. We then showed the necessary condition for $\mathcal{T}$ to be tractable such that the optimal behavior policy can collect data without violating safety constraints. We then proved the first lower bound for this setting under the tractability conditions that scales as $\widetilde{\Omega}(n^{-3/2})$, where $\widetilde{\Omega}$ hides log factors. We then introduced a practical algorithm, SaVeR, that approximates the optimal behavior strategy by computing an upper confidence bound on the variance of the cumulative cost in place of the true cost variances in the optimal behavior strategy. We bound the finite-sample regret (excess MSE) of SaVeR and show that it scales as $\widetilde{O}(n^{-3/2})$ matching the lower bound. Hence, we answer both the questions raised in the introduction positively. In the future, we would like to extend our derivation of optimal data collection strategies and regret analysis of SaVeR to linear/contextual bandits and more general MDPs.

Acknowledgement: J. Hanna was supported in part by American Family Insurance through a research partnership with the University of Wisconsin—Madison's Data Science Institute.

Impact Statement In this paper, we study the safe data collection for policy evaluation in an RL setting under safety constraints. Our paper proposes a new adaptive data collection policy and addresses the theoretical challenges posed by this setting. We focus on algorithmic and theoretical contributions and we do not address the challenges that might stem from incorrect feedback, human bias in feedback, false information, or social disparity in gathering the feedback (or dataset). We therefore leave it to the users who apply our algorithm to use it responsibly and ethically.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Ananye Agarwal, Ashish Kumar, Jitendra Malik, and Deepak Pathak. Legged locomotion in challenging terrains using egocentric vision. *CoRL*, 2022.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9252–9262, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/09a8a8976abcdfdee15128b4cc02f33a-Abstract.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 287–302. Springer, 2008.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300, 2021.

Romain Camilleri, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson. Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214, 2022.

Alexandra Carpentier and Rémi Munos. Finite-time analysis of stratified sampling for monte carlo. In *NIPS-Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.

Alexandra Carpentier and Rémi Munos. Minimax number of strata for online stratified sampling given noisy samples. In *International Conference on Algorithmic Learning Theory*, pages 229–244. Springer, 2012.

Alexandra Carpentier, Remi Munos, and András Antos. Adaptive strategy for stratified monte carlo sampling. *J. Mach. Learn. Res.*, 16:2231–2271, 2015.

Fan Chen, Junyu Zhang, and Zaiwen Wen. A near-optimal primal-dual method for off-policy learning in cmdp. *Advances in Neural Information Processing Systems*, 35: 10521–10532, 2022.

Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric mdps with exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 1855–1863. PMLR, 2021.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

Shutong Ding, Jingya Wang, Yali Du, and Ye Shi. Reduced policy optimization for continuous control with hard constraints. *Advances in Neural Information Processing Systems*, 36, 2024.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. 2014.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Thomas G Fischer. Reinforcement learning in financial markets-a survey. Technical report, FAU Discussion Papers in Economics, 2018.

Xavier Fontaine, Pierre Perrault, Michal Valko, and Vianney Perchet. Online a-optimal design and active linear regression. In *International Conference on Machine Learning*, pages 3374–3383. PMLR, 2021.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3962–3969. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5812.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Shourya Gupta, Utkarsh Suryaman, Rahul Narava, and Shashi Shekhar Jha. Model-based safe reinforcement learning using variable horizon rollouts. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 100–108, 2024.

Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *arXiv preprint arXiv:2112.04553*, 2021.

Josiah P Hanna, Philip S Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In *International Conference on Machine Learning*, pages 1394–1403. PMLR, 2017.

Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. Directional optimism for safe linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2024.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3910–3919, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/bdc4626aa1d1df8e14d80d345b2a442d-Abstract.html.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Shyong Lam and Jon Herlocker. MovieLens Dataset. http://grouplens.org/datasets/movielens/, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Anqi Li, Dipendra Misra, Andrey Kolobov, and Ching-An Cheng. Survival instinct in offline reinforcement learning. *Advances in neural information processing systems*, 36, 2024.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Abhijit Mazumdar, Rafal Wisniewski, and Manuela L Bujorianu. Safe reinforcement learning for constrained markov decision processes with stochastic stopping time. *arXiv preprint arXiv:2403.15928*, 2024.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2020.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

Subhojyoti Mukherjee, Josiah P Hanna, and Robert D Nowak. Revar: Strengthening policy evaluation via reduced variance sampling. In *Uncertainty in Artificial Intelligence*, pages 1413–1422. PMLR, 2022a.

Subhojyoti Mukherjee, Ardhendu S Tripathy, and Robert Nowak. Chernoff sampling for active testing and extension to active regression. In *International Conference on Artificial Intelligence and Statistics*, pages 7384–7432. PMLR, 2022b.

Subhojyoti Mukherjee, Qiaomin Xie, Josiah P Hanna, and Robert Nowak. Speed: Experimental design for policy evaluation in linear heteroscedastic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2962–2970. PMLR, 2024.

Reda Ouhamma, Debabrota Basu, and Odalric Maillard. Bilinear exponential family of mdps: frequentist regret bound with tractable exploration & planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9336–9344, 2023.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Sidney Resnick. *A probability path*. Springer, 2019.

Carlos Riquelme, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active learning for accurate estimation of linear models. In *International Conference on Machine Learning*, pages 2931–2939. PMLR, 2017.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

István Szita. Reinforcement learning in games. *Reinforcement Learning: State-of-the-art*, pages 539–577, 2012.

Tsybakov. Introduction to nonparametric estimation, 2009.

Aaron David Tucker and Thorsten Joachims. Variance-optimal augmentation logging for counterfactual evaluation in contextual bandits. *arXiv preprint arXiv:2202.01721*, 2022.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration for interactive machine learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2887–2897, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4f398cb9d6bc79ae567298335b51ba8a-Abstract.html.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Sharan Vaswani, Lin F Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *arXiv preprint arXiv:2206.06270*, 2022.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. Safe exploration in reinforcement learning: A generalized formulation and algorithms. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. *arXiv preprint arXiv:2202.13234*, 2022.

Tao Wang, Wenbo Du, Chunxiao Jiang, Yumeng Li, and Haijun Zhang. Safety constrained trajectory optimization for completion time minimization for uav communications. *IEEE Internet of Things Journal*, 2024.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

Nuoya Xiong, Yihan Du, and Longbo Huang. Provably safe reinforcement learning with step-wise violation constraints. *Advances in Neural Information Processing Systems*, 36, 2024.

Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, and Simon Shaolei Du. A reduction-based framework for conservative bandits and reinforcement learning. In *International Conference on Learning Representations*, 2021.

Zhaoxing Yang, Haiming Jin, Yao Tang, and Guiyun Fan. Risk-aware constrained reinforcement learning with non-stationary policies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2029–2037, 2024.

Donghao Ying, Yunkai Zhang, Yuhao Ding, Alec Koppel, and Javad Lavaei. Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. *Advances in Neural Information Processing Systems*, 36, 2024.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: a survey. arxiv. *arXiv preprint arXiv:1908.08796*, 2019.

Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. *arXiv preprint arXiv:2401.10700*, 2024.

Rujie Zhong, Duohan Zhang, Lukas Schäfer, Stefano V. Albrecht, and Josiah P. Hanna. Robust On-Policy Sampling for Data-Efficient Policy Evaluation in Reinforcement Learning. In *Proceedings of Neural and Information Processing Systems (NeurIPS)*, 2022. URL http://arxiv.org/abs/2111.14552.

Ruihao Zhu and Branislav Kveton. Safe data collection for offline and online policy learning. *arXiv preprint arXiv:2111.04835*, 2021.

Ruihao Zhu and Branislav Kveton. Safe optimal design with applications in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2436–2447. PMLR, 2022.

A. Appendix

A.1. Related Works

Our work lies at the intersection of two areas: 1) optimal data collection for policy evaluation, and 2) safe sequential decision-making. Optimal data collection for policy evaluation has been studied in reinforcement learning (Antos et al., 2008; Carpentier and Munos, 2012; 2011; Carpentier et al., 2015; Hanna et al., 2017; Mukherjee et al., 2022a; Riquelme et al., 2017; Fontaine et al., 2021; Mukherjee et al., 2024; Zhong et al., 2022) without considering the safety constraints. In the bandit setting the optimal data collection has been studied in the context of estimating a weighted sum of the mean reward associated with each arm. (Antos et al., 2008) study estimating the mean reward of each arm equally well and show that the optimal solution is to pull each arm proportional to the variance of its reward distribution. Since the variances are unknown a priori, they introduce an algorithm that pulls arms in proportion to the empirical variance of each reward distribution. A similar set of works by Carpentier and Munos (2012); Carpentier et al. (2015) extend the above work by introducing a weighting on each arm that is equivalent to the target policy action probabilities in our work. They show that the optimal solution is then to pull each arm proportional to the product of the standard deviation of the reward distribution and the arm weighting. The work of Riquelme et al. (2017); Fontaine et al. (2021); Mukherjee et al. (2024) considers the linear bandit setting to study the policy evaluation setup where actions have different variances. Finally, Mukherjee et al. (2022a) study the policy evaluation setting for tabular MDP. However, these works only look into the policy evaluation setting without considering the safety constraint introduced in (1).

The safe sequential decision-making setup has recently attracted much attention in machine learning (Amodei et al., 2016; Turchetta et al., 2019) and reinforcement learning (Efroni et al., 2020; Wachi and Sui, 2020; Camilleri et al., 2022). In reinforcement learning, and specifically in the bandit setting, safety has been studied in the context of policy improvement. In the bandit literature regret minimization under safety constraints has been studied in Wu et al. (2016); Kazerouni et al. (2017); Amani et al. (2019); Garcelon et al. (2020). In these works the safety requirements are encoded in the form of constraints on the cumulative rewards observed by the learner. These works refer to the setup as conservative bandits because exploration is limited by the constraints on the cumulative reward. The work of Wu et al. (2016) consider the setting of stochastic bandits for policy improvement with a safety constraint similar to (1). However, Kazerouni et al. (2017); Amani et al. (2019); Garcelon et al. (2020); Moradipari et al. (2021); Pacchiano et al. (2021); Hutchinson et al. (2024) study the linear bandit setting under safety constraints where the actions have features associated with them. Note that none of the above works study policy evaluation under safety constraints. Wan et al. (2022); Zhu and Kveton (2021; 2022) analyzes off policy evaluation in the context of designing a non-adaptive policy using inverse probability weighting estimator (as opposed to designing an adaptive policy using certainty equivalence estimator in this work).

In the MDP setting the works of Efroni et al. (2020); Altman (2021); Wachi et al. (2024); Li et al. (2024); Zheng et al. (2024); Xiong et al. (2024); Ding et al. (2024); Wang et al. (2024); Mazumdar et al. (2024) study different variations of the safe exploration in constraint MDPs in both offline and online policy improvement settings. The work of Yang et al. (2024) studies the safe policy improvement in constraint MDP setting under non-stationary policies. The work of Gupta et al. (2024) proposed a safe policy improvement approach for variable horizon setting such that the safe reinforcement learning agent uses a variable look-ahead horizon to avoid unsafe states. The constrained MDP problems have also been looked into from the lens of optimization where Chen et al. (2021; 2022); Qiu et al. (2020); Ding et al. (2020); Vaswani et al. (2022); Ding et al. (2021); Liang et al. (2018); Ying et al. (2024) have proposed a primal-dual sampling-based algorithm to solve CMDPs for the policy improvement setting.

A.2. Previous results and Probability Tools

Proposition 1. ***(Restatement from Carpentier and Munos (2011))*** *In an A-action bandit setting, the estimated return of π after n action-reward samples is denoted by Y_n. Note that the expectation of Y_n after each action has been sampled once is given by V^π. Minimal MSE, $\mathbb{E}_{\mathcal{D}}\left[\left(Y_n - V^\pi\right)^2\right]$, is obtained by taking actions in the proportion:*

$$\mathbf{b}_*(a) \coloneqq \frac{\pi(a)\sigma(a)}{\sum_{a'=1}^{A} \pi(a')\sigma(a')}. \tag{11}$$

where $\mathbf{b}^*(a)$ denotes the optimal sampling proportion.

Lemma A.1. ***(Wald's lemma for variance)*** *(Resnick, 2019) Let $\{\mathcal{F}_t\}$ be a filtration and R_t be a $\mathcal{F}_t$-adapted sequence of i.i.d. random variables with variance σ^2. Assume that $\mathcal{F}_t$ and the σ-algebra generated by $\{R_{t'} : t' \geq t+1\}$ are independent*

and T is a stopping time w.r.t. $\mathcal{F}_t$ with a finite expected value. If $\mathbb{E}\left[R_1^2\right] < \infty$ then

$$\mathbb{E}\left[\left(\sum_{t'=1}^{n} R_{t'} - n\mu\right)^2\right] = \mathbb{E}[n]\sigma^2$$

Lemma A.2. ***(Restatement of Theorem 1 of Mukherjee et al. (2022a))*** *Assume the underlying MDP is an L-depth tree MDP as defined in Definition 3.1. Let the estimated return of the starting state s_1^1 after n state-action-reward samples be defined as $Y_n(s_1^1)$. Let $\mathcal{D}$ be the observed data over n state-action-reward samples. To minimize MSE $\mathbb{E}_{\mathcal{D}}[(Y_n(s_1^1) - V^\pi(s_1^1))^2]$ the optimal sampling proportions for any arbitrary state is given by:*

$$\mathbf{b}_*(a|s_i^\ell) \propto \left(\pi^2(a|s_i^\ell)\left[\sigma^2(s_i^\ell, a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell, a)M^2(s_j^{\ell+1})\right]\right)^{1/2},$$

where, $M(s_j^\ell)$ is the normalization factor defined as follows:

$$M(s_i^\ell) \coloneqq \sum_a \left(\pi^2(a|s_i^\ell)\big(\sigma^2(s_i^\ell, a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell, a)M^2(s_j^{\ell+1})\big)\right)^{1/2}$$

B. Intractable MDP

Proposition 1. *Fix an arbitrary $n > 0$. Then there exists an environment where no algorithm (including the safe oracle $\mathbf{b}_*^k$) can be run that will result in a regret $\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*)$ of $\widetilde{O}(n^{-3/2})$ while satisfying the safety constraint, where $\mathbf{b}_*$ is the unconstrained oracle.*

Proof. We first consider a bandit setting where there are 3 arms, action $\{0\}$ which is the safe action, and actions 1 and 2. Assume $\pi(a) = 1/A$ so that we can ignore its effect on optmal sampling policy $\mathbf{b}_*$

Case 1 (All actions safe): First consider an environment when all actions are safe. That is $\mu^c(0) = 0$ and $\mu^c(1) = 1$ and $\mu^c(2) = 1 - \epsilon$ and reward distributions are bounded between $[0, 1]$. Therefore at round $\ell \in [L]$ we can guarantee for any $\alpha \in (0, 1]$ that

$$\sum_{\ell'=1}^{\ell}\sum_{a=0}^{2} \pi(a)\widehat{\mu}_{c,\ell'}(a) \geq (1-\alpha)\ell \underbrace{\pi_0(0)\mu^c(0)}_{0}, \quad \forall \ell \in [L]$$

where, π_0 always samples safe action 0. Assume a safe oracle that knows the variances of the actions but does not know the means of the actions (both reward and cost means). Therefore from Carpentier and Munos (2011) we know that the optimal way to reduce the MSE $\min_{\mathbf{b}} \mathbb{E}_{\mathcal{D}}[(Y_n^\pi(s_1) - V^\pi(s_1))^2]$ is to run the policy $\mathbf{b}_*(a) \propto \pi(a)\sigma(a)$. We also know from Carpentier and Munos (2011) that there exists an algorithm $\mathcal{A}^{safe}$ (like MC-UCB that tracks $\mathbf{b}_*$) that achieves a regret after n rounds as $\mathcal{R}_n^{\text{safe}} = \widetilde{O}(\frac{K\log(n)}{n^{3/2}})$ where $\widetilde{O}$ hides logarithmic factors and problem dependent factors like $\mathbf{b}_{\min}$.

Case 2 (Some actions are unsafe): In this case, we now analyze a safe oracle algorithm $\mathbf{b}_*^k$. Consider an environment where $\mu^c(0) = 0.5$, $\mu^c(1) = 0.5 + \alpha$, and $\mu^c(2) = 0$. Let the rewards be bounded in $[0, 1]$ again. So action $\{2\}$ is unsafe. Therefore safe oracle policy which first runs action 1 for $C_1 n$ number of times for some $C_1 > 0$. Then it runs the safe action 0 for $C_0 n$ number of times (for some $C_0 > 0$) such that it has enough safety budget and then it runs action 2 for $n(1 - (C_0 + C_1))$ number of times. Let the variance of $\sigma^{r,(2)}(0) = 0.001$, $\sigma^{r,(2)}(1) = 0.001$ and $\sigma^{r,(2)}(2) = 0.25$.

The cost cumulative value over rounds for the algorithm for $\alpha = \frac{1}{4}$ is given by

$$V_{\mathcal{A}}^c = (C_1 n)(0.5 + \alpha) + n(1 - C_0 - C_1)0 + (C_0 n)0.5 = (C_1 n) \cdot \frac{3}{4} + (C_0 n)\frac{2}{4} = \frac{n}{4}\left(3C_1 + 2C_0\right).$$

Then to satisfy the safety budget we have to show that

$$\begin{aligned} & V_{\mathcal{A}}^c \geq n(1-\alpha)0.5 \\ \stackrel{(a)}{\implies} & \frac{n}{4}\left(3C_1 + 2C_0\right) \geq \frac{3n}{8} \\ \implies & 3C_1 + 2C_0 \geq \frac{3}{2} \end{aligned}$$

Say we just want to satisfy the safety constraint, then setting $C_1 = \frac{1}{4}$ and $C_0 = \frac{3}{8}$ in the above equation we can achieve that. Therefore we have that $T_n(1) = \frac{n}{4}$ and $T_n(0) = \frac{3n}{8}$. This implies that $T_n(2) = n - \frac{n}{4} - \frac{3n}{8} = \frac{3n}{8}$. Therefore we get that the loss of $\mathbf{b}_*^k$ is given by

$$\mathcal{L}_n(\pi, \mathbf{b}_*^k) = \sum_{a, T_n(a)>0} \frac{\sigma^{r,(2)}(a)}{T_n(a)} = \frac{8(0.001)^2}{3n} + \frac{4(0.001)^2}{n} + \frac{8(0.25)^2}{3n}$$

Now we calculate the loss of the optimal data collection algorithm following the unconstrained $\mathbf{b}_*$. Note that now $T_n^*(0) = \frac{0.001}{0.001+0.001+0.25} n = \frac{n}{252}$, $T_n^*(1) = \frac{n}{252}$ and $T_n^*(2) = \frac{250n}{252}$. Then the loss of the optimal data collection algorithm following $\mathbf{b}_*$ is given by

$$\mathcal{L}_n^*(\pi, \mathbf{b}_*) = \sum_{a, T_n^*(a)>0} \frac{\sigma^{r,(2)}(a)}{T_n^*(a)} = \frac{252(0.001)^2}{n} + \frac{252(0.001)^2}{n} + \frac{252(0.25)^2}{250n} \approx \frac{2}{4000n} + \frac{15}{n}.$$

It follows then that the regret scales as

$$\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}_*^k) - \mathcal{L}_n^*(\pi, \mathbf{b}_*) = \sum_{a, T_n(a)>0} \frac{\sigma^{r,(2)}(a)}{T_n(a)} - \sum_{a, T_n^*(a)>0} \frac{\sigma^{r,(2)}(a)}{T_n^*(a)} = O\left(\frac{K}{n}\right) \geq \mathcal{R}_n^{\text{safe}} = \widetilde{O}(\frac{K\log(n)}{n^{3/2}}).$$

Note that this regret rate holds for any $C_1 < C_0$ and we cannot shift any more proportion to action $\{2\}$. Therefore the algorithm will choose the sub-optimal safe action $\{0\}$ more than the action that reduces the MSE (to satisfy safety constraint) most resulting in a regret that scales as n^{-1}. So any algorithm (including the safe oracle algorithm) will not be able to achieve the desired regret rate of $\widetilde{O}(n^{-3/2})$. The claim of the proposition follows. □

Remark B.1. **(Tractability condition)** Let $\mathbf{b}$ be any behavior policy that minimizes MSE. However, running $\mathbf{b}$ only once is not enough to guarantee a regret of $\widetilde{O}(n^{-3/2})$. Let $\mathbf{b}$ be run for K_b episodes to guarantee a regret of $\widetilde{O}(n^{-3/2})$. Note that K_b is the number of rounds in the bandit setting. Observe that the number of rounds (or episodes in case of MDP) K_b is behavior policy specific.

Case 1 (Two action bandits): Consider two action bandit setting such that $A = 2$. Further, let $\pi(a) = 1/A$ and the left action has a constraint-value of C_1 while the right action has a constraint-value of C_2. Let the deterministic baseline policy π_0 always choose the left action, while the behavior policy $\mathbf{b}$ chooses the right action. Note that $\mathbf{b}$ may or may not be $\mathbf{b}_*$. Then to satisfy the safety constraint (1) we need that

$$\begin{aligned} (n-K_b)C_1 + K_bC_2 \geq (1-\alpha)nC_1 & \implies nC_1 - K_bC_1 + K_bC_2 \geq nC_1 - \alpha nC_1 \\ & \implies K_b(C_1 - C_2) \leq nC_1\alpha \\ & \implies 1 - \frac{C_2}{C_1} \leq \frac{n\alpha}{K_b} \\ & \implies \frac{K_b}{\alpha}(1 - \frac{C_2}{C_1}) \leq n \\ & \implies n \geq \frac{K_b}{\alpha}\left(1 - \frac{C_2}{C_1}\right) \end{aligned}$$

The above inequality shows two things, (1) the lower bound to the budget n to run the behavior policy $\mathbf{b}$ for K_b rounds and satisfy the safety constraint; (2) The condition $C_1 > C_2$ has to be satisfied so that the RHS is positive.

Case 2 (General multi-armed bandits): Now generalizing this to $A \geq 2$ we can show that the above condition can be modified into

$$\begin{aligned}
&(n-K_b)\mu^c(0) + K_b \min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a) \geq (1-\alpha)n\mu^c(0) \\
\implies & n\mu^c(0) - K_b\mu^c(0) + K_b \min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a) \geq n\mu^c(0) - \alpha n\mu^c(0) \\
\implies & K_b(\mu^c(0) - \min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a)) \leq \alpha n\mu^c(0) \\
\implies & 1 - \frac{\min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a)}{\mu^c(0)} \leq \frac{\alpha n}{K_b} \\
\implies & \frac{K_b}{\alpha}\left(1 - \frac{\min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a)}{\mu^c(0)}\right) \leq n \\
\implies & n \geq \frac{K_b}{\alpha}\left(1 - \frac{\min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a)}{\mu^c(0)}\right)
\end{aligned}$$

The above inequality shows two things, (1) the lower bound to the budget n to run the behavior policy $\mathbf{b}$ for K_b rounds and satisfy the safety constraint for a general K_b armed bandit; (2) The condition $\min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a) < \mu^c(0)$ has to be satisfied so that the RHS is positive.

Case 3 (Tabular MDP): Define $V_c^{\mathbf{b}^-}(s_1)$ as the value of the policy $\mathbf{b}^-$ starting from state s_1. So this policy $\mathbf{b}^-$ can be thought of as the worst possible policy that can be followed by the agent during an episode. Let this policy be run for K_{b^-} episodes. Also, recall that $V_c^{\pi_0}(s_1)$ is the value of the baseline policy π_0 starting from state s_1. It can easily shown following a similar line of argument as case 2 that we need a budget of

$$n \geq \frac{K_{b^-}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right).$$

Again the above inequality shows two things for a general Tree MDP: (1) the lower bound to the budget n to run the behavior policy $\mathbf{b}^-$ for K_{b^-} episodes and satisfy the safety constraint for a Tree MDP; (2) $V_c^{\mathbf{b}^-}(s_1) < V_c^{\pi_0}(s_1)$ so that the RHS is positive.

Now observe that in the first two cases of the bandit setting the $V_c^{\mathbf{b}^-}(s_1)$ yields $\min_{a\in\mathcal{A}\setminus\{0\}} \mu^c(a)$. Therefore combining all three cases we can state the budget $n \geq \frac{K_{b^-}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)$. Now from (Carpentier and Munos, 2012; Mukherjee et al., 2022a) we know that $K_{b^-} \geq C_\sigma(n - \sqrt{n})$ where $C_\sigma \in (0, 1]$ is an MDP dependent parameter that depends on the reward variance of state-action pairs to achieve a regret bound of $\widetilde{O}(n^{-3/2})$. We define the quantity $C_\sigma = \max_{s,a} \frac{\mathbf{b}_*(a|s)}{M(s)}$ where $\mathbf{b}_*(a|s)$ and $M(s)$ are defined in (4) and (5) respectively. Observe that $C_\sigma \in (0, 1)$. Then we have that

$$\begin{aligned}
& n \geq \frac{K_{b^-}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \implies n \geq \frac{C_\sigma(n-\sqrt{n})}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \\
\implies & n \geq \frac{C_\sigma n}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) - \frac{\sqrt{n}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \\
\implies & n\left(1 - \frac{C_\sigma}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)\right) + \frac{\sqrt{n}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \geq 0 \\
\implies & \sqrt{n}\left(\sqrt{n} - \frac{C_\sigma\sqrt{n}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) + \frac{1}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)\right) \geq 0.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\sqrt{n} - \frac{C_\sigma \sqrt{n}}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) + \frac{1}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \geq 0 \\
\implies &\sqrt{n}\left(1 - \frac{C_\sigma}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)\right) \geq -\frac{1}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) \\
\implies &\sqrt{n} \geq \frac{-\frac{1}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)}{\left(1 - \frac{C_\sigma}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)\right)} \\
\implies &\sqrt{n} \geq \frac{\frac{1}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right)}{\frac{C_\sigma}{\alpha}\left(1 - \frac{V_c^{\mathbf{b}^-}(s_1)}{V_c^{\pi_0}(s_1)}\right) - 1}.
\end{aligned}$$

This yields the tractability condition.

C. Tractable MDP and Lower Bounds

Some Definitions for proving Lower Bound: These definitions follow similar definitions in Wagenmaker et al. (2022). Define the Q-function that satisfies the Bellman equation as

$$Q_\ell^\pi(s,a) = R_\ell(s,a) + \sum_{s'} P_\ell\left(s' \mid s,a\right) V_{\ell+1}^\pi\left(s'\right)$$

and $Q_{L+1}^\pi(s,a) = 0$. Define the optimal Q-function as $Q_\ell^{\pi_*}(s,a) \coloneqq \sup_\pi Q_\ell^\pi(s,a), V_\ell^{\pi_*}(s) \coloneqq \sup_\pi V_\ell^\pi(s)$, and let $\pi^\star$ denote an optimal policy. A policy $\widehat{\pi}$ is called ϵ-optimal which satisfies the following

$$V^{\pi_*}(s_1) - V^{\widehat{\pi}}(s_1) \leq \epsilon$$

with probability greater than $1-\delta$ using as few episodes as possible. We further define a few more notations for proving the lower bound. Define the suboptimality gap as

$$\Delta_\ell(s,a) \coloneqq V_\ell^{\pi_*}(s) - Q_\ell^{\pi_*}(s,a).$$

such that $\Delta_\ell(s,a)$ denotes the suboptimality of taking action a in (s,h), and then playing the optimal policy henceforth. Define the state-action visitation distribution as:

$$w_\ell^\pi(s,a) \coloneqq \mathbb{P}_\pi\left[s_\ell = s, a_\ell = a\right], \quad w_\ell^\pi(s) \coloneqq \mathbb{P}_\pi\left[s_\ell = s\right].$$

Note that $w_\ell^\pi(s,a) = \pi_\ell(a|s) w_\ell^\pi(s)$. We denote the maximum reachability of (s,ℓ) by

$$W_\ell(s) \coloneqq \sup_\pi w_\ell^\pi(s).$$

This is the maximum probability with which we could hope to reach (s,ℓ). Define the best-policy gap-visitation complexity as $\mathcal{C}^\star(\mathcal{T})$. Finally, recall that tree MDP is a subset of general MDPs which let us restate the following lemmas on lower bound for unconstrained tree MDPs from Wagenmaker et al. (2022).

Lemma C.1. ***(Divergence Lemma, Restatement of Lemma 4.1 from Wagenmaker et al. (2022))*** *Consider tree MDPs $\mathcal{T}$ and $\mathcal{T}'$ with the same state space $\mathcal{S}$, actions space $\mathcal{A}$, horizon L, and initial state distribution P_0. Fix some $(s,\ell) \in \mathcal{S} \times [L]$, and for any $a \in \mathcal{A}$ let $\nu_\ell(s,a)$ denote the law of the joint distribution of (s', R) where $s' \sim P_{\mathcal{T}}(\cdot \mid s,a)$ and $R \sim R_{\mathcal{T}}(s,a)$. Define the law $\nu'_\ell(s,a)$ analogously with respect to $\mathcal{T}'$. Fix some policy π and let $\mathbb{P}_{\mathcal{T}} = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\mathcal{T}'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on $\mathcal{T}$ and $\mathcal{T}'$ induced by the τ-episode interconnection of π and ν (respectively by π' and ν'). For any almost-sure stopping time τ with respect to filtration $(\mathcal{F}_\tau)$,*

$$\sum_{s,a,h} \mathbb{E}_{\mathcal{T}}\left[N_\ell^\tau(s,a)\right] \mathrm{KL}\left(\nu_\ell(s,a), \nu'_\ell(s,a)\right) \geq \sup_{\xi \in \mathcal{F}_\tau} d\left(\mathbb{P}_{\mathcal{T}}(\xi), \mathbb{P}_{\mathcal{T}'}(\xi)\right)$$

where $d(x,y) = x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ and $N_\ell^\tau(s,a)$ denotes the number of visits to (s,a,ℓ) in the τ episodes.

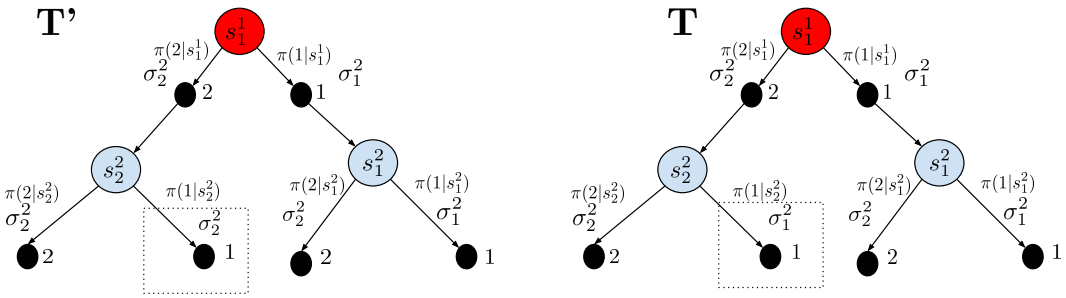


Figure 4. Tractable Tree MDPs $\mathcal{T}$ and $\mathcal{T}'$. The difference between the two Tree MDPs is highlighted in the square box.

Lemma C.2. ***(Proposition 12 from Wagenmaker et al. (2022))*** *Fix some tree MDP $\mathcal{T}$. Then:*

1. *The set of valid state-action visitation distributions on $\mathcal{T}$ is convex.*
2. *For any valid state-action visitation distribution on $\mathcal{T}$, there exists some policy that realizes it.*

Lemma C.3. ***(Restatement of Lemma F.3 from Wagenmaker et al. (2022))*** *In the tree MDP $\mathcal{T}$, fix some $\bar{\ell} \in [L]$. Then*

$$\mathcal{C}^\star(\mathcal{T}) \leq \inf_{\mathbf{b}} \max_{s,a} \frac{1}{w^{\mathbf{b}}_{\bar{\ell}}(s,a)\Delta_\ell(s,a)^2} + \max_{s,\ell} \frac{SAL}{W_\ell(s)}.$$

is the complexity of the Tree MDP $\mathcal{T}$.

Lemma C.4. ***(Proposition 4 from Wagenmaker et al. (2022))*** *The following bounds hold for any unconstrained tree MDP $\mathcal{T}$:*

1. $\mathcal{C}^\star(\mathcal{T}) \leq \frac{L^3 SA}{\epsilon^2}$
2. $\mathcal{C}^\star(\mathcal{T}) \leq \sum_{\ell=1}^{L} \sum_{s,a} \min\left\{\frac{1}{W_\ell(s)\Delta_\ell(s,a)^2}, \frac{W_\ell(s)}{\epsilon^2}\right\} + \frac{L^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$
3. $\mathcal{C}^\star(\mathcal{T}) \leq \sum_{\ell=1}^{L} \sum_{s,a} \frac{1}{\epsilon \max\{\Delta_\ell(s,a),\epsilon\}} + \frac{L^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$

where, $\mathcal{C}^\star(\mathcal{T})$ is the complexity of the Tree MDP $\mathcal{T}$. The second term in $\mathcal{C}^\star(\mathcal{T}), L^2|\mathrm{OPT}(\epsilon)|/\epsilon^2$, captures the complexity of ensuring that after eliminating $\epsilon/W_\ell(s)$-suboptimal actions, sufficient exploration is performed to guarantee the returned policy is ϵ-optimal.

Lemma C.5. ***(Restatement of Theorem 5 in Carpentier and Munos (2012))*** *Let $A \in \mathbb{N}$ be a set of actions for a bandit setting. Let* inf *be the infimum taken over all online sampling algorithms that reduce the MSE and* sup *represent the supremum taken over all environments. Define the regret of the algorithm over the target policy π as $\mathcal{R}_n \coloneqq \mathcal{L}_n(\pi) - \mathcal{L}_n^*(\pi)$ where $\mathcal{L}_n(\pi)$ is the MSE of the target policy following the algorithm. Then:*

$$\inf \sup \mathbb{E}\left[\mathcal{R}_n\right] \geq C\frac{A^{1/3}}{n^{3/2}},$$

where C is a numerical constant, and n is the total budget,

Lemma C.6. *Define the regret of the algorithm over the target policy π as $\mathcal{R}_n \coloneqq \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*)$ where $\mathcal{L}_n(\pi, \mathbf{b})$ is the MSE of the target policy following the algorithm and $\mathbf{b}_*$ is the unconstrained oracle behavior policy. The reward regret in tree MDP $\mathcal{T}$ is lower bounded by*

$$\inf \sup \mathbb{E}\left[\mathcal{R}_n\right] \geq \Omega\left(\frac{\sqrt{SAL^2 \log(1/\delta)}}{n^{3/2}}\right).$$

Proof. We prove this lemma in two steps. In the first step, we prove the minimum number of episodes required by an ϵ-optimal policy $\mathbf{b}$ in tree MDP $\mathcal{T}$ (Figure 4) such that $V^{\mathbf{b}_*}(s_1) - V^{\mathbf{b}}(s_1) \leq \epsilon$. Next in step 2 we show that given this minimum number of episodes, what is the loss suffered by $\mathbf{b}$ against $\mathbf{b}_*$ at the end of episode K.

Step 1 (Minimum episodes): We consider the two tree MDPs $\mathcal{T}$ and $\mathcal{T}'$ shown in Figure 4. We will apply Lemma C.1 on our MDP, $\mathcal{T}$, and MDP $\mathcal{T}'$ which is identical to $\mathcal{T}$ except in state $(s_2^2, 1)$ where we have $\sigma^2(s_2^2, 1) = (\mu - \Delta)(1 - \mu + \Delta)$ in $\mathcal{T}'$ and $\sigma^2(s_1^2, 1) = (\mu + \alpha)(1 - \mu - \alpha)$ for $\mathcal{T}$ and some $\Delta > 0$. This yields a different $\mathbf{b}_*$ for MDP $\mathcal{T}$ than $\mathbf{b}_*$ for $\mathcal{T}'$.

Fix some $\bar{\ell} \in [L]$. Since $\mathcal{T}$ and $\mathcal{T}'$ are identical at all points but this one, we have

$$\sum_{s,a,\ell} \mathbb{E}_{\mathcal{T}}\left[N_\ell^\tau(s,a)\right] \mathrm{KL}\left(\mathrm{Bernoulli}(\mu - \Delta), \mathrm{Bernoulli}(\mu + \alpha)\right)$$
$$= \mathbb{E}_{\mathcal{T}}\left[N_{\bar{\ell}}^\tau(s,a)\right] \mathrm{KL}\left(\mathrm{Bernoulli}(\mu - \Delta), \mathrm{Bernoulli}(\mu + \alpha)\right).$$

where, $\mathbb{E}_{\mathcal{T}}, \mathbb{E}_{\mathcal{T}'}$ denotes the expectation over the data collected in tree MDP $\mathcal{T}$ and $\mathcal{T}'$ respectively following policy $\mathbf{b}_*$.

Let $\mathbf{b}_*$ denote the optimal policy on $\mathcal{T}$, and $\mathbf{b}$ denote the ϵ -optimal policy by any other algorithm. Let the event $\xi = \{\mathbf{b} = \mathbf{b}_*\}$. Since we assume algorithm is δ-correct, and since the optimal policies on $\mathcal{T}$ and $\mathcal{T}'$ differ, we have $\mathbb{P}_{\mathcal{T}}(\xi) \geq 1 - \delta$ and $\mathbb{P}_{\mathcal{T}'}(\xi) \leq \delta$. By Garivier and Kaufmann (2016), we can then lower bound

$$d\left(\mathbb{P}_{\mathcal{T}}(\xi), \mathbb{P}_{\mathcal{T}'}(\xi)\right) \geq \log \frac{1}{2.4\delta}$$

Thus, by Lemma C.1, we have shown that, for any $(s,a), a \neq \mathbf{b}_{*,\bar{\ell}}(s)$,

$$\mathbb{E}_{\mathcal{T}}\left[N_{\bar{\ell}}^\tau(s,a)\right] \geq \frac{1}{\mathrm{KL}\left(\mathrm{Bernoulli}(\mu - \Delta), \mathrm{Bernoulli}(\mu + \alpha)\right)} \cdot \log \frac{1}{2.4\delta}$$

For small $\alpha > 0$, we can bound (see e.g. Lemma 2.7 of Tsybakov (2009))

$$\mathrm{KL}\left(\mathrm{Bernoulli}(\mu - \Delta), \mathrm{Bernoulli}(\mu + \alpha)\right) \leq 6(\Delta - \alpha)^2.$$

Taking $\alpha \to 0$, we have

$$\mathbb{E}_{\mathcal{T}}\left[N_{\bar{\ell}}^\tau(s,a)\right] \geq \frac{1}{6\Delta^2} \cdot \log \frac{1}{2.4\delta}.$$

We can write $\mathbb{E}_{\mathcal{T}}\left[N_{\bar{\ell}}^\tau(s,a)\right] = \mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau w_{\bar{\ell}}^{\mathbf{b}^k}(s,a)\right]$ where $\mathbf{b}^k$ denotes the policy the algorithm played at episode k. Note that all state-visitation distributions lie in a convex set in $[0,1]^{SA}$ and that for any valid state-visitation distribution, there exists some policy that realizes it, by Lemma C.2. By Caratheodory's Theorem, it follows that there exists some set of policies Π with $|\Pi| \leq SA + 1$ such that, for any $\mathbf{b}$ and all $s, a, w_{\bar{\ell}}^{\mathbf{b}}(s,a) = \sum_{\mathbf{b}' \in \Pi} \lambda_{\mathbf{b}'} w_{\bar{\ell}}^{\mathbf{b}'}(s,a)$, for some $\lambda \in \triangle_\Pi$. Note that λ is a distribution over the policies in Π. Letting λ^k denote this distribution satisfying the above inequality for $\mathbf{b}^k$, it follows that

$$\mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau w_{\bar{\ell}}^{\mathbf{b}^k}(s,a)\right] = \mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \sum_{\mathbf{b}\in\Pi} \lambda_{\mathbf{b}}^k w_{\bar{\ell}}^{\mathbf{b}}(s,a)\right]$$
$$= \sum_{\mathbf{b}\in\Pi} \mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \lambda_{\mathbf{b}}^k\right] w_{\bar{\ell}}^{\mathbf{b}}(s,a)$$
$$= \mathbb{E}_{\mathcal{T}}[\tau] \sum_{\mathbf{b}\in\Pi} \frac{\mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \lambda_{\mathbf{b}}^k\right]}{\mathbb{E}_{\mathcal{T}}[\tau]} w_{\bar{\ell}}(s,a).$$

Note that $\sum_{\mathbf{b}\in\Pi} \mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \lambda_{\mathbf{b}}^k\right] = \mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \sum_{\mathbf{b}\in\Pi} \lambda_{\mathbf{b}}^k\right] = \mathbb{E}_{\mathcal{T}}[\tau]$ so it follows that $\left(\frac{\mathbb{E}_{\mathcal{T}}\left[\sum_{k=1}^\tau \lambda_{\mathbf{b}}^k\right]}{\mathbb{E}_{\mathcal{T}}[\tau]}\right)_{\mathbf{b}\in\Pi} \in \triangle_\Pi$. Thus, a δ-correct algorithm must satisfy, for all s, a and some $\lambda \in \triangle_\Pi$,

$$\mathbb{E}_{\mathcal{T}}[\tau] \geq \frac{1}{6\Delta^2 \cdot \sum_{\mathbf{b}\in\Pi} \lambda_{\mathbf{b}} w_{\bar{\ell}}^{\mathbf{b}}(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

Since the set of state visitation distributions is convex, and since for any state-visitation distribution we can find some policy realizing that distribution, for any $\lambda \in \triangle_\Pi$, it follows that there exists some $\mathbf{b}'$ such that, for all s, a, $\sum_{\mathbf{b}\in\Pi} \lambda_{\mathbf{b}} w_\ell^{\mathbf{b}}(s,a) = w_\ell^{\mathbf{b}'}(s,a)$. So, we need, for all s, a

$$\mathbb{E}_{\mathcal{T}}[\tau] \geq \frac{1}{6\Delta^2 \cdot w_\ell^{\mathbf{b}}(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

It follows that every δ-correct algorithm must satisfy

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}[\tau] &\geq \inf_{\mathbf{b}} \max_{s,a} \frac{1}{6\Delta^2 \cdot w_\ell^{\mathbf{b}}(s,a)} \cdot \log \frac{1}{2.4\delta}, \\ &\gtrsim \mathcal{C}^\star(\mathcal{T}) \cdot \log \frac{1}{2.4\delta} - \max_{s,\ell} \frac{SAL}{W_\ell(s)}\end{aligned}$$

from which the first inequality follows, and the second inequality follows from Lemma C.3.

The second term in $\mathcal{C}^\star(\mathcal{T}), L^2|\mathrm{OPT}(\epsilon)|/\epsilon^2$, captures the complexity of ensuring that after eliminating $\epsilon/W_\ell(s)$-suboptimal actions, sufficient exploration is performed to guarantee the returned policy is ϵ-optimal. Using Lemma C.4 we have that $\mathcal{C}^\star(\mathcal{T}), L^2|\mathrm{OPT}(\epsilon)|/\epsilon^2$ will be no worse than L^3SA/ϵ^2, it could be much better, if in the MDP the number of (s,a,ℓ) with $\Delta_\ell(s,a) \lesssim \epsilon/W_\ell(s)$ is small (note that since $\Delta_\ell(s,a) \geq \Delta_{\min}(s,\ell)$ by definition, $\mathrm{OPT}(\epsilon)$ will only contain states for which the minimum non-zero gap is less than $\epsilon/W_\ell(s)$). Wagenmaker et al. (2022) obtains the bounds on $\mathcal{C}^\star(\mathcal{T})$ in Lemma C.4, providing an interpretation of $\mathcal{C}^\star(\mathcal{T})$ in terms of the maximum reachability, and illustrating $\mathcal{C}^\star(\mathcal{T})$ is no larger than the minimax optimal complexity. This implies that

$$\mathbb{E}_{\mathcal{T}}[\tau] \gtrsim \Omega\left(\frac{SAL^2}{\epsilon^2}\log(1/\delta)\right).$$

Hence the $V^{\mathbf{b}^K}(s_1^1) - V^{\mathbf{b}_*}(s_1^1) \leq \epsilon$ for $K \geq \dfrac{SAL^2}{\epsilon^2}\log(1/\delta)$.

Step 2 (Bound regret in $\mathcal{T}$): In the $\mathcal{T}$ in Figure 4 we now have

$$M(s_1^1) = \sqrt{2\sigma_1^2 + \sigma_2^2} + \sqrt{2\sigma_2^2 + \sigma_1^2}, \quad M(s_1^2) = M(s_2^2) = \sigma_1 + \sigma_2$$

Define confidence interval $\beta_L^K = L\sqrt{SA\log(SAL^2/\delta)/n}$. It can be shown using pointwise uncertainty estimation from Corollary 3 that

$$|\widehat{\sigma}_{K,1} - \sigma_1| \leq \beta_L^K, \quad |\widehat{\sigma}_{K,2} - \sigma_2| \leq \beta_L^K \tag{12}$$

holds with probability greater than $1-\delta$, where the $\widehat{\sigma}_{K,1}, \widehat{\sigma}_{K,2}$ denote the estimated variances after K episodes. Then the loss of the agnostic algorithm at the end of the K-th episode is given by

$$\begin{aligned}\mathcal{L}_n^K(\pi, \mathbf{b}) &= \frac{\sqrt{2\widehat{\sigma}_{K,1}^2 + \widehat{\sigma}_{K,2}^2} + \sqrt{2\widehat{\sigma}_{K,2}^2 + \widehat{\sigma}_{K,1}^2}}{n} \\ &\overset{(a)}{\geq} \frac{\sqrt{2(\sigma_1^2 - \beta_L^K) + \sigma_2^2 - \beta_L^K} + \sqrt{2(\sigma_2^2 - \beta_L^K) + \sigma_1^2 - \beta_L^K}}{n} \\ &= \frac{\sqrt{2\sigma_1^2 + \sigma_2^2 - 3\beta_L^K} + \sqrt{2\sigma_2^2 + \sigma_1^2 - 3\beta_L^K}}{n} \\ &\overset{(b)}{\geq} \frac{\sqrt{2\sigma_1^2 + \sigma_2^2} + \sqrt{2\sigma_2^2 + \sigma_1^2}}{n} - C\frac{\beta_L^K}{n}\end{aligned}$$

where, (a) follows from concentration inequality in (12), and (b) follows for some appropriate constant $C > 0$. Then for

$K \geq \frac{SAL^2}{\epsilon^2}\log(1/\delta)$ (from step 1) we have the total loss as

$$\mathcal{L}_n(\pi, \mathbf{b}) = \mathcal{L}_n^K(\pi, \mathbf{b}) \geq \left(\frac{\sqrt{2\sigma_1^2+\sigma_2^2}+\sqrt{2\sigma_2^2+\sigma_1^2}}{n} - \frac{\beta_L^K}{n}\right)\frac{SAL^2}{\epsilon^2}\log(1/\delta)$$
$$\overset{(a)}{\geq} \underbrace{\frac{\sqrt{2\sigma_1^2+\sigma_2^2}+\sqrt{2\sigma_2^2+\sigma_1^2}}{n}}_{\mathcal{L}_n(\pi,\mathbf{b}_*)} + \frac{\beta_L^K}{n}$$
$$\mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*) \overset{(b)}{\geq} \frac{\sqrt{SAL^2\log(SAL^2/\delta)}}{n\sqrt{K}} = \Omega\left(\frac{\sqrt{SAL^2\log(1/\delta)}}{n^{3/2}}\right)$$

where, (a) follows by first setting $\epsilon = 1/\sqrt{n}$ and then noting that

$$\left(\sqrt{2\sigma_1^2+\sigma_2^2}+\sqrt{2\sigma_2^2+\sigma_1^2} - \beta_L^K\right)(SAL^2)\log(1/\delta) \geq \frac{\sqrt{2\sigma_1^2+\sigma_2^2}+\sqrt{2\sigma_2^2+\sigma_1^2}}{n} + \frac{\beta_L^K}{n}.$$

Also note that $\mathcal{L}_n(\pi, \mathbf{b}_*) = \frac{\sqrt{2\sigma_1^2+\sigma_2^2}+\sqrt{2\sigma_2^2+\sigma_1^2}}{n}$. The (b) follows by substituting the value of β_L^K. The claim of the lemma follows. □

Theorem 1. ***(Lower Bound, formal)*** *Let $\pi(a|s) = \frac{1}{A}$ for each state $s \in \mathcal{S}$. Assume the MDP $\mathcal{M}$ is tractable under Assumption 3.2 and satisfies (7). Then the reward regret is lower bounded by*

$$\mathbb{E}\left[\mathcal{R}_n\right] = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*^k) \geq \begin{cases} \Omega\left(\max\left\{\frac{A^{1/3}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 A^{2/3}}{n^{3/2}}\right)\right\}\right), & \textbf{\textit{(MAB)}} \\ \Omega\left(\max\left\{\frac{\sqrt{SAL^2}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 SAL^2}{n^{3/2}}\right)\right\}\right) & \textbf{\textit{(Tabular MDP)}} \end{cases}$$

where, $\Delta_0 = |V_c^{\mathbf{b}_*^k}(s_1^1) - V_c^{\pi_0}(s_1^1)|$ and $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)$ is the hardness parameter.

Proof. We follow a reduction-based proof technique to prove this lower bound (Yang et al., 2021).

Step 1 (Reduction): First recall we have that the regret for any online algorithm **Alg** that minimizes the MSE $\mathcal{L}_n(\pi)$ is given by $\mathcal{R}_n(\textbf{Alg}) = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*(\pi, \mathbf{b}_*^k)$, where $\mathcal{L}_n^*(\pi, \mathbf{b})$ is the MSE of the oracle algorithm. We also assume $\pi(a) = 1/A$ for all $a \in \mathcal{A}$, and $\sigma(a) \geq \frac{1}{16}$ for all a.

Now consider any sequential decision-making problem $\mathfrak{A}$ (for instance a multi-armed bandit problem) such that there exists $\xi \in \mathbb{R}$ (a constant solely depending on the sequential decision-making problem, e.g., the number of actions in bandits, or state-action-horizon in tabular RL), an instance of problem $\mathfrak{A}$ where for the budget n large enough and any algorithm **Alg** we have from Lemma C.5 and Lemma C.6 that:

$$\mathbb{E}\left[\mathcal{R}_n^{\mathfrak{A}}(\textbf{Alg})\right] \geq \frac{\xi}{n^{3/2}}, \tag{13}$$

For instance, in the MAB case $\xi = A^{1/3}$ with A the number of arms and in tabular RL $\xi = SAL^2$. Using this non-conservative (unconstraint) lower bound, we show our lower bound for the safe setting for the problem $\mathfrak{A}$ with a baseline policy π_0. We assume the MDP $\mathcal{T} \subset \mathcal{M}$ where we run the behavior policy $\mathbf{b}_*^k$ satisfies Assumption 3.2. This is required because otherwise we will not be able to run the behavior policy a sufficient number of times to reach a regret bound of $\widetilde{O}(n^{-3/2})$(see Proposition 1). To do so, let's consider any safe algorithm (that is to say it satisfies safety constraint) noted as $\textbf{Alg}_c$. We assume this algorithms selects behavior policies $(\mathbf{b}^t)_{t\in[n]}$ and let $\mathcal{N}_0$ denotes the set of episodes in $\{1, \ldots, K\}$ where $\textbf{Alg}_c$ selects the safe policy π_0. Let $|\mathcal{N}_0| = N_0$ and $\Delta_0 \coloneqq |V^{\mathbf{b}_*^k} - V_c^{\pi_0}|$. Here we assume the budget n is large such

that $n \geq SAL^2/\epsilon^2$ for some $\epsilon > 0$ (see Lemma C.6) and

$$
\begin{aligned}
n &\geq \sqrt{\frac{\xi}{\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)} + \frac{\xi^2}{4\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}} \\
\implies n^2 &\geq \frac{\xi}{\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)} + \frac{\xi^2}{4\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}. \\
\implies n &\geq \frac{\xi}{n\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)} + \frac{\xi^2}{4n\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}
\end{aligned}
$$

Step 2 (Loss estimate): Let $\mathcal{L}(N_0)$ be the loss suffered in first N_0 episodes. We now distinguish two cases:

(a) If $\mathbb{E}\left[\mathcal{L}(N_0)\right] \geq \frac{\xi}{n\alpha V_c^{\pi_0}(s_1^1) \cdot \left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)}$, then the definition of the regret implies that:

$$
\mathbb{E}\left[\mathcal{R}_n^{\mathfrak{A}}(\mathbf{Alg})\right] = \mathbb{E}\left[\mathcal{L}(N_0)\right] \cdot \Delta_0 \geq \frac{\xi \Delta_0}{n\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)}. \tag{14}
$$

(b) If $\mathbb{E}\left[\mathcal{L}(N_0)\right] < \frac{\xi}{n\alpha V^{\pi 0} \cdot \left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)}$, then let's note $\mathcal{N}_0^C = \left\{i_1, i_2, \cdots, i_{|\mathcal{N}_0^c|}\right\}$ the set of episodes where $\mathbf{Alg}_c$ does not execute the baseline policy π_0. Now consider the safety budget (similar to Definition 1 of Yang et al. (2021)) we have:

$$
\begin{aligned}
B_{\mathcal{N}_0^c}\left(\mathbf{Alg}_c\right) &= \max_{t \in \mathcal{N}_0^c} \mathbb{E} \sum_{k=1}^{t} \left[(1-\alpha) V_c^{\pi_0}(s_1^1) - V^{\pi^t}(s_1^1)\right] \\
&= \max_{t \in \mathcal{N}_0^c} \mathbb{E} \sum_{k=1}^{t} \left[V^{\mathbf{b}_*^k}(s_1^1) - V^{\pi^t}(s_1^1) - \alpha V_c^{\pi_0}(s_1^1) - \left(V_c^{\mathbf{b}_*^k}(s_1^1) - V_c^{\pi_0}(s_1^1)\right)\right] \\
&= \max_{t \in \mathcal{N}_0^c} \mathbb{E}\left[\mathcal{R}_{\mathcal{N}_0^C}^{\mathfrak{A}}\left(\mathcal{A}_c\right)(t)\right] - \left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right) t,
\end{aligned}
$$

where $\Delta_0 = V_c^{\mathbf{b}_*^k}(s_1^1) - V_c^{\pi_0}(s_1^1)$ is the difference between the constraint value of the optimal policy and the baseline policy and $\mathbb{E}\left[R_{\mathfrak{A}}^{\mathcal{N}_0^C}\left(\mathcal{A}_c\right)(t)\right]$ is the regret incurred by the episodes $\{i_k\}_{k \in [t]}$. Therefore, for any $t \in [|\mathcal{T}_0^c|]$, by (13) we have that there exists an instance u (for instance in a bandit problem u is the means of each arm) of $\mathfrak{A}$ such that $\mathbb{E}\left[R_{\mathfrak{A}}^{\mathcal{N}_0^C}\left(\mathcal{A}_c\right)(t)\right] \geq \frac{\xi}{t^{3/2}}$. Let $t_0 = \frac{\xi^2}{4n\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}$, then there exists an instance such that

$$
\begin{aligned}
B_{\mathcal{N}_0^C}\left(\mathbf{Alg}_c\right) \geq \frac{\xi}{t_0^{3/2}} - \left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right) t_0 &= \frac{4\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^3 n^{3/2}}{\xi^2} - \frac{\xi^2}{4(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)} \frac{1}{n^2} \\
&\overset{(a)}{\geq} \frac{(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)^2 \xi^2}{n^{3/2}}.
\end{aligned}
$$

where, (a) follows as $n^{3/2} - n^{-2} \geq n^{-3/2}$. Combining the safety condition in Equation (1), we have

$$
\mathbb{E}\left[\mathcal{L}(\mathcal{N}_0)\right] \geq \frac{B_{\mathcal{N}_0}\left(\mathbf{Alg}_c\right)}{\alpha V_c^{\pi_0}(s_1^1)} \gtrsim \frac{\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2 \xi^2}{\alpha V_c^{\pi_0}(s_1^1) n^{3/2}}.
$$

By the same derivation of Equation (14), we have

$$
\mathbb{E}\left[R_n^{\mathfrak{A}}(\mathbf{Alg})\right] \gtrsim \frac{\xi^2 \Delta_0}{n\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)} \overset{(a)}{\geq} \frac{\xi^2}{n^{3/2}\alpha V_c^{\pi_0}(s_1^1) \cdot (\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)}. \tag{15}
$$

where, (a) follows for $\Delta_0 \geq 1/\sqrt{n}$. Combining Equation (13), 14, and 15 we can show that

$$
\mathbb{E}\left[R_n^{\mathfrak{A}}(\mathbf{Alg})\right] \gtrsim \max\left\{\frac{\xi}{n^{3/2}}, \frac{(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)^2 \xi^2}{(\alpha V_c^{\pi_0}(s_1^1))^2 n^{3/2}}\right\}.
$$

Step 3 (Combine with MAB:) Now considering that safe oracle $\mathbf{b}_*^k$ is also an online algorithm **Alg**, we can drop the notation. Then for multi-armed bandits, by Lemma C.5, we choose $\xi = A^{1/3}$. Then we have

$$\mathbb{E}\left[\mathcal{R}_n\right] \gtrsim \max\left\{\frac{A^{1/3}}{n^{3/2}}, \frac{\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}{(\alpha V_c^{\pi_0}(s_1^1))^2}\left(\frac{A^{2/3}}{n^{3/2}}\right)\right\} \overset{(a)}{=} \min\left\{\frac{A^{1/3}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 A^{2/3}}{n^{3/2}}\right)\right\}.$$

where, (a) follows from the problem complexity parameter $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)$ when $\pi(a) = 1/A$ and $\sigma(a) \geq 1/16$ for the bandit setting.

Step 4 (Combine with tabular RL:) For tabular RL, by Lemma C.6, we choose $\xi = \sqrt{SAL^2}$. Then we have

$$\mathbb{E}\left[\mathcal{R}_n\right] \gtrsim \max\left\{\frac{\sqrt{SAL^2}}{n^{3/2}}, \frac{\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)^2}{(\alpha V_c^{\pi_0}(s_1^1))^2}\left(\frac{SAL^2}{n^{3/2}}\right)\right\} \overset{(a)}{=} \min\left\{\frac{\sqrt{SAL^2}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 SAL^2}{n^{3/2}}\right)\right\}.$$

where, (a) follows from the problem complexity parameter $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0)$ when $\pi(a) = 1/A$ and $\sigma(a) \geq 1/16$. This concludes the proof. □

Remark C.7. **(Comparing regret)** Observe that the regret lower bound is proved on $\mathcal{R}'_n = \mathcal{L}_n(\pi) - \mathcal{L}_n^*(\pi)$ which assumes that we can exactly solve for the oracle sampling solution. However, $\overline{\mathcal{L}}_n^*(\pi)$ in $\mathcal{R}_n$ is an upper bound to $\mathcal{L}_n^*(\pi)$ and so we cannot directly compare $\mathcal{R}_n$ with $\mathcal{R}'_n$. However, since $\mathcal{R}'_n$ gives a lower bound by directly solving for the oracle solution, we conjecture that this is the lower bound to $\mathcal{R}_n$. Proving this conjecture we leave it to future works.

D. Proof of Tree Agnostic MSE

Theorem 2. ***(formal)*** *Let Assumption 3.2 hold. Then the MSE of the SaVeR for* $\frac{n}{\log(SAn(n+1)/\delta)} \geq 32(LSA^2)^2 + \frac{SA}{\min_{s,a}\Delta^{c,(2)}(s,a)} + \frac{1}{4H_{*,(2)}^2}$ *is bounded by*

$$\mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k) \leq \frac{M^2(s_1^1)}{n} + \frac{8AM^2(s_1^1)}{n^2} + \frac{16A^2M^2(s_1^1)}{n^3} + \frac{M^2(s_1^1)}{n}\left(32MLSA + H_{*,(2)}\right)^2 + 2\sum_{t=1}^{n}\frac{2\eta + 4\eta^2}{n^2}$$
$$+ O\left(\frac{(2\eta + 4\eta^2)(LSA^2)^2 H_{*,(2)}^2 M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s) n^{3/2}}\right)$$

with probability $(1-\delta)$. *The* $M = \sum_{\ell=1}^{L}\sum_{s_j^\ell} M(s_j^\ell)$, *and* $H_{*,(2)} = \sum_{\ell=1}^{L}\sum_{s_j^\ell} H_{*,(2)}(s_j^\ell)$ *is the problem complexity parameter. The total predicted constraint violations is bounded by*

$$\mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2 + O\left(\frac{(2\eta + 4\eta^2)(LSA^2)^2 H_{*,(2)}^2 M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s) n^{1/2}}\right)$$

with probability $(1-\delta)$, *where* $M_{\min} \coloneqq \min_s M(s)$.

Proof. **Step 1 (Sampling rule):** First note that agnostic SaVeR samples by the following rule

$$\text{Play } \mathbf{b}^k = \begin{cases} \pi_x & \text{if } \widehat{Z}^{k-1} \geq 0, k \leq \sqrt{K} \\ \widehat{\mathbf{b}}^k & \text{if } \widehat{Z}^{k-1} \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}^{k-1} < 0 \end{cases} \tag{16}$$

where, $\widehat{Z}_L^{k-1} \coloneqq \sum_{k'=1}^{k-1}(Y_{c,L}^{\mathbf{b}^{k'}}(s_1^1) - \beta_L^{k'}(s,a)) - (1-\alpha)(k-1)V_c^{\pi_0}(s_1^1)$ is the safety budget till the k-th episode.

Step 2 (MSE Decomposition): Now recall that the agnostic algorithm does not know the variances and the means. We define the good cost event when the oracle has a good estimate of the cost mean. This is stated as follows:

$$\xi_{c,K} \coloneqq \bigcap_{\substack{1\leq k\leq K,\\ 1\leq a\leq A, 1\leq s\leq S}} \left\{\left|\widehat{\mu}_{c,L}^k(s,a) - \mu^c(s,a)\right| \leq (2\eta + 4\eta^2)L\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^k(s,a)}}\right\} \tag{17}$$

where, $n = KL$ and K is the number of episodes and L is the length of horizon of each episode. The exploration policy π_x results in a good constraint estimate of state-action tuples. This is shown in Corollary 4. We define the good variance event as

$$\xi_{v,K} := \bigcap_{\substack{1\le k\le K,\\ 1\le a\le A, 1\le s\le S}} \left\{ |\widehat{\sigma}_L^k(s,a) - \sigma(s,a)| \le (2\eta + 4\eta^2) L \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^k(s,a)}} \right\}. \tag{18}$$

We define the safety budget event

$$\xi_{Z,K} := \bigcap_{1\le k\le K} \left\{ \widehat{Z}^k \ge 0 \right\}. \tag{19}$$

Using the definition of MSE, and Lemma A.1 we can show that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}\right] \\
&\le \sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1,a)}{\underline{T}_L^{(2),K}(s_1^1,a)}\right] \mathbb{E}[T_L^K(s_1^1,a)\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}] \\
&\quad + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{s_j^2} P(s_j^2|s_1^1,a)\mathbf{Var}[Y_n(s_j^2)]\mathbb{E}[T_L^K(s_j^2,a)\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}] \\
&\le \sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1,a)}{\underline{T}_L^{(2),K}(s_1^1,a)}\right] \mathbb{E}[T_L^K(s_1^1,a)\mathbb{I}\{\xi_{Z,K} \cap \mathbb{I}\{\xi_{v,K}\}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
&\quad + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^{L} \sum_{s_j^\ell} P(s_j^\ell|s_1^1,a) \sum_{a'} \pi^2(a'|s_j^\ell)\left[\frac{\sigma^2(s_j^\ell,a')}{\underline{T}_L^{(2),K}(s_j^\ell,a')}\right] \mathbb{E}[T_L^K(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}]
\end{aligned} \tag{20}$$

which implies that SaVeR does not need to know the reward means $\mu(s,a)$. Hence, the MSE of SaVeR is bounded by

$$\begin{aligned}
\mathcal{L}_n(\pi) &\le \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \xi_{v,K}\right]}_{\textbf{Part A, } \widehat{Z}_n \ge 0\textbf{, safety event holds}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}^C\}\right]}_{\textbf{Part B, } \widehat{Z}_n < 0\textbf{, constraint violation}} \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{v,K}^C\}\right]}_{\textbf{Part D, Variance event does not hold}} \\
&\le \sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1,a)}{\underline{T}_L^{(2),K}(s_1^1,a)}\right] \mathbb{E}[T_L^K(s_1^1,a)\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}] \\
&\quad + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^{L} \sum_{s_j^\ell} P(s_j^\ell|s_1^1,a) \sum_{a'} \pi^2(a'|s_j^\ell)\left[\frac{\sigma^2(s_j^\ell,a')}{\underline{T}_L^{(2),K}(s_j^\ell,a')}\right] \mathbb{E}[T_L^K(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}^C\}\right]}_{\textbf{Part B, } \widehat{Z}_n < 0\textbf{, constraint violation}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{v,K}^C\}\right]}_{\textbf{Part D, Variance event does not hold}}.
\end{aligned}$$

Divide the total budget n into two parts, n_f when $\sum_{j=1}^k \mathbb{I}\{\widehat{Z}^j \ge 0\}$ is true, then $\mathbf{b}_*$ or π_x is run. Hence define

$$n_f := \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{s_j^\ell} \sum_{a'=1}^{A} \mathbb{E}[T_\ell^k(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}].$$

The other part consist of $n_u = n - n_f$ number of samples when $\sum_{j=1}^{k} \mathbb{I}\{\widehat{Z}^k < 0\}$ and only π_0 is run. Hence we define,

$$n_u = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{s_j^\ell} \sum_{a'=1}^{A} \mathbb{E}[T_\ell^k(s_j^\ell, a') \mathbb{I}\{\xi_{Z,K}^C\}].$$

Step 3 (Sampling of SaVeR for $\widehat{Z}^k \geq 0$): First note that when $\widehat{Z}^k \geq 0$ the SaVeR samples at episode k and round $\ell+1$ the action $\arg\max_a U_{\ell+1}^k(s_i^{\ell+1}, a)$ where

$$U_\ell^k(s_i^\ell, a) := \frac{\widehat{\mathbf{b}}_\ell^k(a|s_i^\ell)}{T_\ell^k(s_i^\ell, a)} \leq \frac{\pi(a|s_i^\ell)}{T_\ell^k(s_i^\ell, a)} \Bigg(\sigma(s_i^\ell, a) + (2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_\ell^k(s_i^\ell, a)}} + \underbrace{\gamma^2 \sum_{a'} \pi(a'|s_i^\ell) \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell|a') \widehat{M}(s_j^{\ell+1})}_{\mathbf{B}(s_i^\ell)} \Bigg). \tag{21}$$

Let $\ell + 1 > 2SA$ be the time at which a given state-action (s_i^ℓ, p') is visited for the last time, i.e., $T_\ell^k(p') = T_L^K(p') - 1$ and $T_{\ell+1}^k(p') = T_L^K(p')$. Note that as $n = KL \geq 4SA$, there is at least one state-action pair (s_i^ℓ, p') such that this happens, i.e. such that it is visited after the initialization phase. Note that under Assumption 3.2 it is possible to visit each (s, a) atleast once. Since the SaVeR chooses to visit (s_i^ℓ, p') at time $\ell + 1$, we have for any state-action pair (s_i^ℓ, p')

$$U_{\ell+1}^k(s_i^{\ell+1}, p) \leq U_{\ell+1}^k(s_i^{\ell+1}, p'). \tag{22}$$

From (21) and using the fact that $T_\ell^k(s_i^\ell, p') = T_L^K(s_i^\ell, p') - 1$, we can show that

$$\begin{aligned} U_{\ell+1}^k(s_i^{\ell+1}, p') &\leq \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_t^k(s_i^{\ell+1}, p')} \left((2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_t^k(s_i^{\ell+1}, p') - 1}} + \mathbf{B}(s_i^{\ell+1}) \right) \\ &= \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_L^K(s_i^{\ell+1}, p') - 1} \left((2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^K(s_i^{\ell+1}, p') - 1}} + \mathbf{B}(s_i^{\ell+1}) \right). \end{aligned} \tag{23}$$

Also note that

$$U_{\ell+1}^k(s_i^{\ell+1}, p) = \frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_t^k(s_i^{\ell+1}, p)} \left((2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_t^k(s_i^{\ell+1}, p) - 1}} + B(s_i^{\ell+1}) \right) \overset{(a)}{\geq} \frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_L^K(s_i^{\ell+1}, p)}. \tag{24}$$

where, (a) follows as $T_t(p) \leq T_L^K(p, s_i^{\ell+1})$ (i.e., the number of times p has been visited can only increase after time ℓ). Combining (22), (23), (24) we can show that for any action p:

$$\frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_L^K(p, s_i^{\ell+1})} \leq \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_L^K(p', s_i^{\ell+1}) - 1} \left((2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^K(s_i^{\ell+1}, p') - 1}} + \mathbf{B}(s_i^{\ell+1}) \right). \tag{25}$$

Note that in the above equation, there is no dependency on ℓ, and thus, the probability that (25) holds for any $(s_i^{\ell+1}, p)$ and for any $(s_i^{\ell+1}, p')$ such that action $(s_i^{\ell+1}, p')$ is visited after the initialization phase, i.e., such that $T_L^K(s_i^{\ell+1}, p') > 2$ depends on the probability of event $\xi_{Z,n}$.

Step 4. ((Lower bound on $T_L^K(s_i^\ell, p)$ for $\widehat{Z}^k \geq 0$): If a state-action tuple (s_i^ℓ, p) is less visited compared to its optimal allocation without taking into account the initialization phase, i.e., $T_L^K(s_i^\ell, p) - 2 < \mathbf{b}(p|s_i^\ell)(n - 2A)$, then from the constraint $\sum_{p'} \left(T_L^K(s, p') - 2\right) = n - 2SA$ and the definition of the optimal allocation, we deduce that there exist at least another state-action tuple s_i^ℓ, p' that is over-visited compared to its optimal allocation without taking into account the initialization phase, i.e., $T_L^K(s_i^\ell, p') - 2 > \mathbf{b}(s_i^\ell, p')(n - 2A)$. Note that for this action, $T_L^K(s_i^\ell, p') - 2 > \mathbf{b}_*(p'|s_i^\ell)(n - 2SA) \geq 0$, so we know that this specific action is taken at least once after the initialization phase and that it satisfies (25). Recall that we have defined $M(s_i^\ell) = \sum_a \pi(a|s_i^\ell)\sigma(s_i^\ell, a)$. Further define $M = \sum_{\ell=1}^{L} \sum_{s_i^\ell} M(s_i^\ell)$. Using the definition of the optimal

allocation $T_L^{*,K}(s_i^\ell, p') = n_f \frac{\mathbf{b}_*(p'|s_i^\ell)}{M(s_i^\ell)}$, and the fact that $T_L^K(s_i^\ell, p') \geq \mathbf{b}_*(p'|s_i^\ell)(n_f - 2SA) + 2$, (25) may be written as for any state-action tuple (s_i^ℓ, p)

$$\begin{aligned}\frac{\mathbf{b}_*(p|s_i^\ell)}{T_L^K(s_i^\ell, p)} &\leq \frac{\mathbf{b}_*(p'|s_i^\ell)}{T_L^{*,K}(p', s_i^\ell)} \frac{n_f}{(n_f - 2SA)} \left((2\eta + 4\eta^2) \sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^K(s_i^{\ell+1}, p') - 1}} + \mathbf{B}(s_i^{\ell+1}) \right) \\ &\leq \frac{M(s_i^\ell)}{n_f} + \frac{4SAM(s_i^\ell)}{n_f^2} + \frac{(2\eta + 4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(s_i^\ell) n_f^{3/2}} \end{aligned} \tag{26}$$

because $n_f \geq 4SA$. By rearranging (26), we obtain the lower bound on $T_L^K(s_i^\ell, p)$:

$$\begin{aligned} T_L^K(s_i^\ell, p) &\geq \frac{\mathbf{b}_*(p|s_i^\ell)}{\frac{M(s_i^\ell)}{n_f} + \frac{4SAM(s_i^\ell)}{n_f^2} + \frac{(2\eta + 4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(s_i^\ell) n_f^{3/2}}} \\ &\overset{(a)}{\geq} T_L^{*,K}(s_i^\ell, p) - \frac{(2\eta + 4\eta^2)\mathbf{b}_*(p|s_i^\ell)\sqrt{\log(SAn(n+1)/\delta)}}{M(s_i^\ell)\mathbf{b}_{*,\min}^{3/2}(s_i^\ell) n_f^{3/2}} - 4A\mathbf{b}_*(p|s_i^\ell), \end{aligned} \tag{27}$$

where in (a) we use $1/(1+x) \geq 1 - x$ (for $x > -1$). Note that the lower bound holds on $\xi_{c,K}$ for any state-action (s_i^ℓ, p).

Step 5. (Upper bound on $T_L^K(s_i^\ell, p)$ for $\widehat{Z}^k \geq 0$): Now using (27) and the fact that n_f is given by $\sum_{\ell=1}^L \sum_{s_j^\ell} \sum_{a'=1}^A \mathbb{E}[T_L^K(s_j^\ell, a') \mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\} \cap \mathbb{I}\{\xi_{v,K}\}] = n_f$, we obtain

$$\begin{aligned} T_L^K(s_i^\ell, p) = n_f - \sum_{p' \neq p} T_L^K(s_i^\ell, p') &\leq \left(n_f - \sum_{p' \neq p} T_L^{*,K}(s_i^\ell, p') \right) \\ &+ \sum_{p' \neq p} \left(\frac{(2\eta + 4\eta^2)\mathbf{b}_*(p'|s_i^\ell)\sqrt{\log(SAn(n+1)/\delta)}}{M(s_i^\ell)\mathbf{b}_{*,\min}^{3/2}(s_i^\ell) n_f^{3/2}} + 4A\mathbf{b}_*(p'|s_i^\ell) \right). \end{aligned}$$

Now since $\sum_{p' \neq p} \mathbf{b}_*(p'|s_i^\ell) \leq 1$ we can show that

$$T_L^K(s_i^\ell, p) \leq T_L^{*,K}(s_i^\ell, p) + \frac{(2\eta + 4\eta^2)\mathbf{b}_*(p|s_i^\ell)\sqrt{\log(SAn(n+1)/\delta)}}{M(s_i^\ell)\mathbf{b}_{*,\min}^{3/2}(s_i^\ell) n_f^{3/2}} + 4A. \tag{28}$$

Step 6 (Bound part A): We now bound the part A using (26)

$$
\begin{aligned}
&\sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1,a)}{\underline{T}_L^{(2),K}(s_1^1,a)}\right]\mathbb{E}[T_L^K(s_1^1,a)\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}\cap\mathbb{I}\{\xi_{v,K}\}]\\
&\quad+\gamma^2\sum_a\pi^2(a|s_1^1)\sum_{\ell=2}^{L}\sum_{s_j^\ell}P(s_j^\ell|s_1^1,a)\sum_{a'}\pi^2(a'|s_j^\ell)\left[\frac{\sigma^2(s_j^\ell,a')}{\underline{T}_L^{(2),K}(s_j^\ell,a')}\right]\mathbb{E}[T_L^K(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}]\\
&\overset{(a)}{\leq}\left(\frac{M(s_1^1)}{n_f}+\frac{4SAM(s_1^1)}{n_f^2}+\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(p|s_i^\ell)n_f^{3/2}}\right)^2 n_f\\
&\quad+\gamma^2\sum_a\pi^2(a|s_1^1)\sum_{\ell=2}^{L}\sum_{s_j^\ell}P(s_j^\ell|s_1^1,a)\left(\frac{M(s_j^\ell)}{n_f}+\frac{4SAM(s_j^\ell)}{n_f^2}+\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(p|s_i^\ell)n_f^{3/2}}\right)^2 n_f\\
&=\frac{M^2(s_1^1)}{n_f}+\frac{8AM^2(s_1^1)}{n_f^2}+\frac{16A^2M^2(s_1^1)}{n_f^3}+O\left(\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(p|s_i^\ell)n_f^{3/2}}\right)\\
&\quad+\gamma^2\sum_a\pi^2(a|s_1^1)\sum_{\ell=2}^{L}\sum_{s_j^\ell}P(s_j^\ell|s_1^1,a)\left(\frac{M^2(s_j^\ell)}{n_f}+\frac{8AM^2(s_j^\ell)}{n_f^2}+\frac{16A^2M^2(s_j^\ell)}{n_f^3}\right.\\
&\quad\left.+O\left(\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\mathbf{b}_{*,\min}^{3/2}(p|s_j^\ell)n_f^{3/2}}\right)\right)
\end{aligned}
$$

where, in (a) follows from the definition of $M(s)$ and n_f.

Step 7 (Upper Bound to Constraint Violation): In this step we bound the quantity $\mathcal{C}_n(\pi)=\sum_{j=1}^k\mathbb{I}\{\widehat{Z}^j<0,\mathbf{b}^j\in\{\widehat{\mathbf{b}}^k,\pi_0\}\}$. Define the number of times the policy $\mathbf{b}_*$ is played till episode k is $T^k(\mathbf{b}_*)$ and the number of times the baseline policy is played is given by $T^k(\pi_0)$. Observe that $\mathcal{C}_n(\pi)=\sum_{j=1}^k\mathbb{I}\{\widehat{Z}^j<0,\mathbf{b}^j\in\{\widehat{\mathbf{b}}^k,\pi_0\}\}=T^K(\pi_0)\mathbb{I}\{\xi_{Z,K}^C\}$ as when the constraint are violated policy π_0 is sampled. Let $\tau=\max\left\{k\leq K\text{ and }n_f\geq\frac{\log(SAn(n+1)/\delta)}{\min_{s,a}\Delta^{c,\alpha,(2)}(s,a)}\mid\mathbf{b}^k=\pi_0\right\}$ be the last episode in which the baseline policy is played. We will define formally the gap $\Delta^{c,\alpha,(2)}(s,a)$ later. Observe that the constraint violation can be re-stated as follows:

$$
\begin{aligned}
&\sum_{k=1}^{\tau}Y_{\mathbf{b}^k}^c(s_1^1)\coloneqq\sum_{k=1}^{\tau}\sum_a\mathbf{b}^k(a|s_1^1)\left(\widehat{\mu}_L^{c,k}(s_1,a)+\sum_{s_j^2}P(s_j^2|s_1^1,a)Y_{\mathbf{b}^k}^c(s_j^2)\right)<(1-\alpha)\tau V_{\pi_0}^c(s_1^1)\\
&\implies\sum_{k=1}^{\tau}\sum_a\mathbf{b}^k(a|s_1^1)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a)+\sum_{s_j^2}P(s_j^2|s_1^1,a)\underline{Y}_{\mathbf{b}^k}^c(s_j^2)\right)<(1-\alpha)\tau V_{\pi_0}^c(s_1^1)\\
&\overset{(a)}{\implies}\sum_{k=1}^{\tau}\sum_a\mathbf{b}^k(a|s_1^1)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a)+\sum_{s_j^2}P(s_j^2|s_1^1,a)\underline{Y}_{\mathbf{b}^k}^c(s_j^2)\right)\\
&\qquad<(1-\alpha)\sum_{k=1}^{\tau}\pi_0(0|s_1^1)\left(\mu^c(s_1^1,0)+\sum_{s_j^2}P(s_j^2|s_1^1,0)V_{\pi_0}^c(s_j^2)\right)
\end{aligned}
$$

$$\implies \sum_{k=1}^{\tau}\sum_{a} T_L^k(s_1^1,a)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a)+\sum_{s_j^2} P(s_j^2|s_1^1,a)\underline{Y}^c_{\mathbf{b}^k}(s_j^2)\right)$$

$$< (1-\alpha)\sum_{k=1}^{\tau} T_L^k(s_1^1,a)\left(\mu^c(s_1^1,0)+\sum_{s_j^2} P(s_j^2|s_1^1,0)V^c_{\pi_0}(s_j^2)\right)$$

$$\overset{(b)}{\implies} \underbrace{\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a)}_{\textbf{Part A}} + \sum_a T_L^\tau(s_1^1,a)\sum_{s_j^2} P(s_j^2|s_1^1,a)\underline{Y}^c_{\mathbf{b}^k}(s_j^2)$$

$$< \underbrace{(1-\alpha)\sum_a T_L^\tau(s_1^1,0)\mu^c(s_1^1,0)}_{\textbf{Part B}} + (1-\alpha)T_L^\tau(s_1^1,0)\sum_{s_j^2} P(s_j^2|s_1^1,0)V^c_{\pi_0}(s_j^2). \tag{29}$$

Comparing **Part A** and **Part B** for level $\ell = 1$ we observe that the constraint violation must satisfy

$$\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a) < (1-\alpha)T_L^\tau(s_1^1,0)\mu^c(s_1^1,0)$$

which can be reduced as follows

$$T_L^{\tau-1}(s_1^1,0) \le \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1+\sum_{a=1}^{A} N(s_1^1,a)\right).$$

where $\Delta^{c,\alpha}(s_1^1,a) \coloneqq (1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a)$ and

$$\begin{aligned} N(s_1^1,a) &\coloneqq T_L^{\tau-1}(s_1^1,a)\cdot\left((1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)/T_L^{\tau-1}(s_1^1,a)}\right) \\ &= \Delta^{c,\alpha}(s_1^1,a)T_L^{\tau-1}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)T_L^{\tau-1}(s_1^1,a)} \end{aligned} \tag{30}$$

is a bound on the decrease in $\widehat{Z}_\tau$ in the first $\tau-1$ rounds due to choosing action a in s_1^1. We will now bound $N(s_1^1,a)$ for each a. Now observe

$$\begin{aligned} \Delta^{c,\alpha}(s_1^1,a) = (1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a) &= \mu^c(s_1^1,0) - \alpha\mu^c(s_1^1,0) - \mu^c(s_1^1,a) \\ &= -(\mu^{*,c}(s_1^1) - \mu^c(s_1^1,0)) - \alpha\mu^c(s_1^1,0) + (\mu^{*,c}(s_1^1) - \mu^c(s_1^1,a)) \\ &= -\Delta^c(s_1^1,0) - \alpha\mu^c(s_1^1,0) + \Delta^c(s_1^1,a). \end{aligned}$$

where, $\mu^{*,c}(s_1^1) = \max_a \mu^c(s_1^1,a)$. Let $J(n_f) = \frac{(2\eta+4\eta^2)\mathbf{b}_*(p|s_i^\ell)\sqrt{\log(SAn(n+1)/\delta)}}{M(s_i^\ell)\mathbf{b}_{*,\min}^{3/2}(s_i^\ell)n_f^{3/2}}$. The first case is $\Delta^{c,\alpha}(s_1^1,a) > 0$, i.e. $\Delta^c(s_1^1,a) > \Delta^c(0) + \alpha\mu^c(0)$. These are the unsafe actions as $\Delta^{c,\alpha}(s_1^1,a) \coloneqq (1-\alpha)\mu^c(0) - \mu^c(s_1^1,a) > 0$ we have from (28)

$$T_n(s_1^1,a) \le T_n^*(s_1^1,a) + J(n_f) + 4A = \frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f + J(n_f) + 4A$$

Plugging this back in $N(s_1^1,a)$ we get

$$\begin{aligned} &N(s_1^1,a) = \Delta^{c,\alpha}(s_1^1,a)T_{\tau-1}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)T_{\tau-1}(s_1^1,a)} + J(n_f) \\ &\le \frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f\Delta^{c,\alpha}(s_1^1,a) + 4A\Delta^{c,\alpha}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f + 4A\right)} + J(n_f) \\ &\overset{(a)}{\le} \frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f\Delta^{c,\alpha}(s_1^1,a) + 4A\Delta^{c,\alpha}(s_1^1,a) + c_1\sqrt{\Delta^{c,\alpha,(2)}(s_1^1,a)\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f + 4A\right)} + J(n_f) \\ &\le 2\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}n_f\Delta^{c,\alpha}(s_1^1,a) + 4A\Delta^{c,\alpha}(s_1^1,a)\right) + J(n_f). \end{aligned} \tag{31}$$

where, (a) follows for $n_f \geq \frac{\log(SAn(n+1)/\delta)}{\min_a \Delta^{c,\alpha,(2)}(s_1^1,a)}$. The other case is $\Delta^{c,\alpha}(s_1^1,a) < 0$, i.e. $\Delta^c(s_1^1,a) < \Delta^c(s_1^1,0) + \alpha\mu^c(s_1^1,0)$ then only safe actions are pulled. Then

$$\begin{aligned}
N(s_1^1,a) &\leq -\Delta^{c,\alpha}(s_1^1,a)T_{\tau-1}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)T_{\tau-1}(s_1^1,a)} + J(n_f) \\
&= \underbrace{-\Delta^{c,\alpha}(s_1^1,a)}_{a} T_{\tau-1}(s_1^1,a) + \underbrace{c_1\sqrt{\log(An(n+1)/\delta)}}_{b}\sqrt{T_{\tau-1}(s_1^1,a)} + J(n_f) \\
&\overset{(a)}{\leq} -\frac{\log(An(n+1)/\delta)}{4\Delta^{c,\alpha}(s_1^1,a)} = \frac{\log(An(n+1)/\delta)}{4(\Delta^c(0) + \alpha\mu^c(0) - \Delta^c(s_1^1,a))} \\
&\overset{(b)}{\leq} 4\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M} n_f(\Delta^c(0) + \alpha\mu^c(0) - \Delta^c(s_1^1,a))\right)
\end{aligned} \tag{32}$$

where, (a) follows by using $ax^2 + bx \leq -b^2/4a$ for $a < 0$, and (b) follows as $n_f \geq \frac{\log(An(n+1)/\delta)}{\min^+_{a\in\mathcal{A}\setminus\{0\}} \pi(s_1^1,a)\sigma(s_1^1,a)\Delta^{c,\alpha,(2)}(s_1^1,a)}$ which implies

$$\begin{aligned}
&\frac{\log(An(n+1)/\delta)}{n_f} \leq 4\left(\sum_{\mathcal{A}\setminus\{0\}} \pi(s_1^1,a)\sigma(s_1^1,a) \overset{+}{\min}\{\Delta^c(s_1^1,a), \Delta^c(0) - \Delta^c(s_1^1,a)\}\right)^2 \\
&\implies \frac{\log(An(n+1)/\delta)}{\sum_{\mathcal{A}\setminus\{0\}} \pi(s_1^1,a)\sigma(s_1^1,a) \min^+\{\Delta^c(s_1^1,a), \Delta^c(0) - \Delta^c(s_1^1,a)\}} \\
&\leq 4\left(\sum_{\mathcal{A}\setminus\{0\}} \pi(s_1^1,a)\sigma(s_1^1,a) \overset{+}{\min}\{\Delta^c(s_1^1,a), \Delta^c(0) - \Delta^c(s_1^1,a)\}\right) n_f.
\end{aligned}$$

Plugging everything back in (32), we get

$$\begin{aligned}
n_u = T_{\tau-1}(s_1^1,0) &= \frac{1}{\alpha\mu^c(0)}\left(\sum_{a=1}^{A} N(s_1^1,a)\right) \\
&\leq \frac{2}{\alpha\mu^c(0)} \sum_{a\in\mathcal{A}_u} \Delta^c(s_1^1,a)\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M} n_f\right) + \frac{4}{\alpha\mu^c(0)} \sum_{a\in\mathcal{A}_s\setminus\{0\}} \left(\Delta^c(0) - \Delta^c(s_1^1,a)\right)\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M} n_f\right) \\
&= \frac{6}{\alpha\mu^c(0)} \sum_{a\in\mathcal{A}\setminus\{0\}} \overset{+}{\min}\{\Delta^c(s_1^1,a), \Delta^c(0) - \Delta^c(s_1^1,a)\}\left(\frac{\pi(s_1^1,a)\sigma(s_1^1,a)}{M}(n - n_u)\right) \leq \frac{H_{*,(2)}}{2}\frac{n}{M}.
\end{aligned} \tag{33}$$

It follows then that for the state s_1^1

$$n_u(s_1^1) \leq \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1 + \sum_{a=1}^{A} N(s_1^1,a)\right) \leq \frac{H_{*,(2)}(s_1^1)}{2}\frac{n}{M(s_1^1)}$$

where

$$\begin{aligned}
H_{*,(2)}(s_i^\ell) &\coloneqq \sum_a \mathbf{b}_*(a|s_i^\ell) \overset{+}{\min}\{\Delta^c(s_i^\ell,a), \Delta^c(s_i^\ell,0) - \Delta^c(s_i^\ell,a)\}, \\
M(s_i^\ell) &\coloneqq \sum_a \sqrt{\pi^2(a|s_i^\ell)\left(\sigma^2(s_i^\ell,a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell,a)M^2(s_j^{\ell+1})\right)}.
\end{aligned} \tag{34}$$

For an arbitrary level $\ell \in [L]$, we can show using (29) that the constraint violation must satisfy

$$\begin{aligned}
&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_{a} T_L^{\tau}(s_i^{\ell'},a)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}} T_L^{\tau}(s_i^{\ell'},0)\mu_0^c(s_i^{\ell'},0)\\
\overset{(a)}{\Longrightarrow}&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_{a}\left(T_L^{*,K}(s_i^{\ell'},a) - 4A\mathbf{b}_*(a|s_i^{\ell'}) - O\left(\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\min_{s_i^{\ell'}}\mathbf{b}_{*,\min}^{k,(3/2)}(s_i^{\ell'})n_f^{3/2}}\right)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a)\\
&\quad< (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0) + 4A + O\left(\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\min_{s_i^{\ell'}}\mathbf{b}_{*,\min}^{k,(3/2)}(s_i^{\ell'})n_f^{3/2}}\right)\right)\mu^c(s_i^{\ell'},0)\\
\Longrightarrow&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_{a}\left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right)\mu^c(s_i^{\ell'},0)\\
&\quad+8LSA^2(\mu^c(s_i^{\ell'},0) + \underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a)) + O\left(\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\min_{s_i^{\ell'}}\mathbf{b}_{*,\min}^{k,(3/2)}(s_i^{\ell'})n_f^{3/2}}\right)\\
\Longrightarrow&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_{a}\left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right)\mu^c(s_i^{\ell'},0)\\
&\quad+8LSA^2(\mu_{0,L}^{c,\tau}(s_i^{\ell'},a) + \widehat{\mu}_L^{c,\tau}(s_i^{\ell'},a) - \sqrt{\frac{\log((SAn(n+1)/\delta)}{2T_L^{\tau}(s_i^{\ell'},a)}})\\
&\quad+O\left(\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\frac{(2\eta+4\eta^2)\sqrt{\log(SAn(n+1)/\delta)}}{\min_{s_i^{\ell'}}\mathbf{b}_{*,\min}^{k,(3/2)}(s_i^{\ell'})n_f^{3/2}}\right)\\
\Longrightarrow&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_{a}\left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\max_s\mu^c(s,0)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right) + 16LSA^2\\
&\quad+O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s\mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right)
\end{aligned} \tag{35}$$

where, (a) follows as $\mu(s,a) \in (0,1]$ for all s,a and using (27) and (28). Summing over all states s_j^{ℓ} till level L we can show that

$$\begin{aligned}
n_u = \sum_{\ell=1}^{L}\sum_{s_j^{\ell}} T_L^{*,K}(s_j^{\ell},0) &\le \frac{n}{2}\sum_{\ell=1}^{L}\sum_{s_j^{\ell}}\frac{H_{*,(2)}(s_j^{\ell})}{M(s_j^{\ell})} + 16LSA^2 + O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s\mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right)n_f\\
&\overset{(a)}{\le}\frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2 + O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s\mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{1/2}}\right)\\
&\overset{(b)}{\le}\frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2 + O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s\mathbf{b}_{*,\min}^{k,(3/2)}(s)n^{1/2}}\right)
\end{aligned} \tag{36}$$

where, in (a) we define $M_{\min} = \min_s M(s)$, and $H_{*,(2)} = \sum_{\ell=1}^{L}\sum_{s_j^{\ell}} H_{*,(2)}(s_j^{\ell})$, and (b) follows by setting $n_f = n - n_u$. Finally, observe that $16LSA^2$ does not depend on the episode K, and the quantity $O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s\mathbf{b}_{*,\min}^{k,(3/2)}(s)n^{1/2}}\right)$ decreases with n.

Step 8 (Lower Bound to Constraint Violation): For the lower bound to the constraint we equate Equation (29) to 0 and

show that

$$\underbrace{\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a)}_{\textbf{Part A}} + \sum_a T_L^\tau(s_1^1,a)\sum_{s_j^2} P(s_j^2|s_1^1,a)\underline{Y}^c_{\mathbf{b}^k}(s_j^2)$$
$$= \underbrace{(1-\alpha)\sum_a T_L^\tau(s_1^1,0)\mu^c(s_1^1,0)}_{\textbf{Part B}} + (1-\alpha)T_L^\tau(s_1^1,0)\sum_{s_j^2} P(s_j^2|s_1^1,0)V^c_{\pi_0}(s_j^2).$$

Again comparing **Part A** and **Part B** for level $\ell = 1$ we observe that the lower bound to constraint violation must satisfy

$$\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a) = (1-\alpha)T_L^\tau(s_1^1,0)\mu^{c,}(s_1^1,0)$$

which can be reduced as

$$\sum_a T_L^{\tau-1}(s_1^1,0) \geq \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1+\sum_{a=1}^A \underline{N}(s_1^1,a)\right).$$

where $\Delta^{c,\alpha}(s_1^1,a) \coloneqq (1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a)$ and

$$\begin{aligned}\underline{N}(s_1^1,a) &\coloneqq T_L^{\tau-1}(s_1^1,a)\cdot\left((1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)/T_L^{\tau-1}(s_1^1,a)}\right)\\ &= \Delta^{c,\alpha}(s_1^1,a)T_L^{\tau-1}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)T_L^{\tau-1}(s_1^1,a)}\\ &\overset{(a)}{\geq} \Delta^{c,\alpha}(s_1^1,a)\left(T_L^{*,K}(s_1^1,a) - 4A\mathbf{b}_*(a|s_1^1)\right) + c_1\sqrt{\log(An(n+1)/\delta)\left(T_L^{*,K}(s_1^1,a) - 4A\mathbf{b}_*(a|s_1^1)\right)}\end{aligned}$$

where, (a) follows from (27). Then we can show that

$$T_L^{\tau-1}(s_1^1,0) \geq \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1+\sum_{a=1}^A \underline{N}(s_1^1,a)\right) \geq \frac{n_f}{M(s_1^1)}\left(\frac{H_{*,(2)}(s_1^1)}{8} - \frac{A}{2}\frac{H_{*,(2)}(s_1^1)}{M(s_1^1)}\right) - 16SA$$

Similarly for any arbitrary level $\ell \in [L]$ following the same way as step 7 above it can be shown that

$$\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a\left(T_L^{*,K}(s_i^{\ell'},a) + 4A\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) \geq (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0) - 4A\mathbf{b}_*(0|s_i^{\ell'})\right)\mu^c(s_i^{\ell'},0)$$
$$\implies \sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a T_L^{*,K}(s_j^\ell,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_j^\ell,a) \geq (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}} T_L^{*,K}(s_i^\ell,0)\mu^c(s_i^\ell,0) - 16LSA^2.$$

For the state s_j^ℓ we can show that

$$\begin{aligned}\sum_{\ell'=1}^{\ell}\sum_{s_j^{\ell'}} T_L^{*,K}(s_j^{\ell'},0) &\geq \frac{1}{\alpha\max_s\mu^c(s_j^\ell,0)}\left(1+\sum_{\ell'=1}^{\ell}\sum_{s_j^{\ell'}}\sum_{a=1}^A \underline{N}(s_j^{\ell'},a)\right)\\ &\geq \sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}\left(\frac{H_{*,(2)}(s_j^\ell)}{8} - \frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right) - 16LSA^2 - O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}^{k,(3/2)}_{*,\min}(s)n_f^{3/2}}\right).\end{aligned}$$

Finally summing over all states s_j^ℓ and level L we can show that

$$\begin{aligned}\sum_{\ell=1}^{L}\sum_{s_j^\ell} T_L^{*,K}(s_j^\ell,0) \geq \sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}&\left(\frac{H_{*,(2)}(s_j^\ell)}{8} - \frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)\\ &- 16LSA^2 - O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}^{k,(3/2)}_{*,\min}(s)n_f^{3/2}}\right).\end{aligned} \tag{37}$$

Again, observe that $16LSA^2$ does not depend on the episode K.

Step 9 (Bound Part B): Then from (37) we can show that

$$
\begin{aligned}
&\frac{M(s_1^1)}{\sum_{\ell=1}^{L}\sum_{s_j^\ell} T_L^{*,K}(s_j^\ell,0)} \\
&\le \frac{M(s_1^1)}{\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}-\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)-16LSA^2-O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right)} \\
&\overset{(a)}{\le} (M(s_1^1)+16LSA^2)\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{M(s_j^\ell)}{n_f}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}+\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)+O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right) \\
&\le (M(s_1^1)+16LSA^2)\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{M(s_j^\ell)}{n_f}\left(2+H_{*,(2)}(s_j^\ell)\right)+O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right) \\
&\overset{(b)}{\le} (M(s_1^1)+16LSA^2)\frac{M}{n_f}\left(2+H_{*,(2)}\right)+O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right)
\end{aligned}
$$

where, (a) follows for $1/(x-c)\le x+c$ for $x^2\ge 1+c^2$ and $c>0$. The (b) follows for $H_{*,(2)}=\sum_{\ell=1}^{L}\sum_{s_j^\ell}H_{*,(2)}(s_j^\ell)$ and $M=\sum_{\ell=1}^{L}\sum_{s_j^\ell}M(s_j^\ell)$. It follows then by setting $n_f=n-n_u$ that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{Z,K}^C\}\right] \\
&\overset{(a)}{\le}\left(\frac{(M(s_1^1)+16LSA^2)M\left(2+H_{*,(2)}\right)}{n_f}+O\left(\frac{(2\eta+4\eta^2)L\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n_f^{3/2}}\right)\right)^2 n_u \\
&\overset{(b)}{\le}\frac{(M(s_1^1)+16LSA^2)^2 n_u}{(n-n_u)^2}\left(2+H_{*,(2)}\right)^2+O\left(\frac{(2\eta+4\eta^2)L^2S^2A^4H_{*,(2)}^2M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)(n-n_u)^{3/2}}\right) \\
&\overset{(c)}{\le}\frac{(M(s_1^1)+16LSA^2)^2 H_{*,(2)}n}{(n-H_{*,(2)}n)^2}\left(2+H_{*,(2)}\right)^2+O\left(\frac{(2\eta+4\eta^2)L^2S^2A^4H_{*,(2)}^2M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)(n-H_{*,(2)}n)^{3/2}}\right) \\
&\le\frac{M^2(s_1^1)}{n}\left(32MLSA^2+H_{*,(2)}\right)^2+O\left(\frac{(2\eta+4\eta^2)L^2S^2A^4H_{*,(2)}^2M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n^{3/2}}\right)
\end{aligned}
$$

where, (a) follows from Lemma A.1, (b) follows from definition of $H_{*,(2)}$, and (c) follows from (36).

Step 10 (Combine everything): Combining everything from step 5, step 8 and setting $\delta=1/n^2$ we can show that the MSE of SaVeRscales as

$$
\begin{aligned}
\mathcal{L}_n(\pi,\widehat{\mathbf{b}}^k) &\le \frac{M^2(s_1^1)}{n}+\frac{8AM^2(s_1^1)}{n^2}+\frac{16A^2M^2(s_1^1)}{n^3}+\frac{M^2(s_1^1)}{n}\left(32MLSA+H_{*,(2)}\right)^2 \\
&\quad+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{v,K}^C\}\right]}_{\textbf{Part D, Variance event does not hold}} \\
&\overset{(a)}{\le}\frac{M^2(s_1^1)}{n}+\frac{8AM^2(s_1^1)}{n^2}+\frac{16A^2M^2(s_1^1)}{n^3}+\frac{M^2(s_1^1)}{n}\left(32MLSA+H_{*,(2)}\right)^2+2\sum_{t=1}^{n}\frac{2\eta+4\eta^2}{n^2} \\
&\quad+O\left(\frac{(2\eta+4\eta^2)L^2S^2A^4H_{*,(2)}^2M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}_{*,\min}^{k,(3/2)}(s)n^{3/2}}\right)
\end{aligned}
\tag{38}
$$

where, (a) follows as $\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{c,K}^C\}\right]\le 2\eta+4\eta^2$ and using the low error probability of the cost event from Lemma F.4 and variance event from Corollary 3. The claim of the theorem follows.

□

E. Proof of Tree Oracle MSE

Proposition 2. ***(formal)*** *Let Assumption 3.2 hold. Then the MSE of the oracle for* $\frac{n}{\log(SAn(n+1)/\delta)} \geq 32(LSA^2)^2 + \frac{SA}{\min_{s,a}\Delta^{c,(2)}(s,a)} + \frac{1}{4H^2_{*,(2)}}$ *is bounded by*

$$\mathcal{L}_n(\pi, \mathbf{b}^k_*) \leq \frac{M^2(s^1_1)}{n} + \frac{8AM^2(s^1_1)}{n^2} + \frac{16A^2M^2(s^1_1)}{n^3} + \frac{M^2(s^1_1)}{n}\left(32MLSA^2 + H_{*,(2)}\right)^2 + 2\sum_{t=1}^{n}\frac{2\eta+4\eta^2}{n^2} + \frac{2}{n}$$

with probability $(1-\delta)$*. The* $M = \sum_{\ell=1}^{L}\sum_{s^\ell_j} M(s^\ell_j)$*, and* $H_{*,(2)} = \sum_{\ell=1}^{L}\sum_{s^\ell_j} H_{*,(2)}(s^\ell_j)$ *is the problem complexity parameter. The total predicted constraint violations is bounded by*

$$\mathcal{C}^*_n(\pi, \mathbf{b}^k_*) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2$$

with probability $(1-\delta)$*, where* $M_{\min} \coloneqq \min_s M(s)$*.*

Proof. **Step 1 (Sampling rule):** We follow the proof technique of Theorem 2. Note that the oracle tree algorithm knows the variances of reward and constraints values (but does not know the mean of either) and samples by the following rule

$$\mathbf{b}^k_* = \begin{cases} \pi_x, & \text{if } \widehat{Z}^{k-1}_L \geq 0, k \leq \sqrt{K} \\ \mathbf{b}_*, & \text{if } \widehat{Z}^{k-1}_L \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}^{k-1}_L < 0 \end{cases}. \tag{39}$$

where, $\widehat{Z}^{k-1}_L \coloneqq \sum_{k'=1}^{k-1}(Y^{\mathbf{b}^{k'}}_{c,L}(s^1_1) - \beta^{k'}_L(s,a)) - (1-\alpha)(k-1)V^{\pi_0}_c(s^1_1)$ is the safety budget till the k-th episode.

Step 2 (MSE Decomposition): Now recall that the oracle knows the variances but does not know the means (constraint and reward). We define the good constraint event when the oracle has a good estimate of the constraint mean. This is stated as follows:

$$\xi_{c,K} \coloneqq \bigcap_{\substack{1\leq k\leq K,\\ 1\leq a\leq A, 1\leq s\leq S}} \left\{\left|\widehat{\mu}^{c,k}_L(s,a) - \mu^c(s,a)\right| \leq (2\eta+4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T^k_L(s,a)}}\right\} \tag{40}$$

where, $n = KL$ and K is the number of episodes and L is the length of horizon of each episode. Define $c_1 = 2\eta + 4\eta^2$.

The exploration policy π_e results in a good constraint estimate of state-action tuples. This is shown in Corollary 4.

We also define the safety budget event $\xi_{Z,K} \coloneqq \bigcap_{1\leq k\leq K}\left\{\widehat{Z}^k \geq 0\right\}$. Now using Lemma A.1 we can show that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s^1_1) - V_\pi(s^1_1)\right)^2 \mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}\right] \leq \sum_a \pi^2(a|s^1_1)\left[\frac{\sigma^2(s^1_1,a)}{\underline{T}^{(2),K}_L(s^1_1,a)}\right]\mathbb{E}[T^K_L(s^1_1,a)\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}] \\
&\quad + \gamma^2\sum_a \pi^2(a|s^1_1)\sum_{s^2_j} P(s^2_j|s^1_1,a)\mathbf{Var}[Y_n(s^2_j)]\mathbb{E}[T^K_L(s^2_j,a)\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}] \\
&\leq \sum_a \pi^2(a|s^1_1)\left[\frac{\sigma^2(s^1_1,a)}{\underline{T}^{(2),K}_L(s^1_1,a)}\right]\mathbb{E}[T^K_L(s^1_1,a)\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}] \\
&\quad + \gamma^2\sum_a \pi^2(a|s^1_1)\sum_{\ell=2}^{L}\sum_{s^\ell_j} P(s^\ell_j|s^1_1,a)\sum_{a'}\pi^2(a'|s^\ell_j)\left[\frac{\sigma^2(s^\ell_j,a')}{\underline{T}^{(2),K}_L(s^\ell_j,a')}\right]\mathbb{E}[T^K_L(s^\ell_j,a')\mathbb{I}\{\xi_{Z,K}\}\cap\mathbb{I}\{\xi_{c,K}\}]
\end{aligned}$$

which implies that the oracle does not need to know the reward means $\mu(a)$. Hence, Using the definition of MSE we can show that the MSE of oracle is bounded by

$$\begin{aligned}
\mathcal{L}_n(\pi) \leq & \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}\right]}_{\textbf{Part A, } \widehat{Z}_n \geq 0\textbf{, safety event holds}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}^C\}\right]}_{\textbf{Part B, } \widehat{Z}_n < 0\textbf{, constraint violation}} \\
& + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}} \\
& \leq \sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1,a)}{\underline{T}_L^{(2),K}(s_1^1,a)}\right]\mathbb{E}[T_L^K(s_1^1,a)\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
& + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^{L} \sum_{s_j^\ell} P(s_j^\ell|s_1^1,a) \sum_{a'} \pi^2(a'|s_j^\ell)\left[\frac{\sigma^2(s_j^\ell,a')}{\underline{T}_L^{(2),K}(s_j^\ell,a')}\right]\mathbb{E}[T_L^K(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
& + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{Z,K}^C\}\right]}_{\textbf{Part B, } \widehat{Z}_n < 0\textbf{, constraint violation}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}}
\end{aligned}$$

Divide the total budget n into two parts, n_f when $\sum_{j=1}^k \mathbb{I}\{\widehat{Z}^j \geq 0\}$ is true, then $\mathbf{b}_*$ is run. Hence define

$$n_f \coloneqq \sum_{k=1}^{K}\sum_{\ell=1}^{L}\sum_{s_j^\ell}\sum_{a'=1}^{A} \mathbb{E}[T_\ell^k(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}].$$

The other part consist of $n_u = n - n_f$ number of samples when $\sum_{j=1}^k \mathbb{I}\{\widehat{Z}^k < 0\}$ and only π_0 is run. Hence we define,

$$n_u = \sum_{k=1}^{K}\sum_{\ell=1}^{L}\sum_{s_j^\ell}\sum_{a'=1}^{A} \mathbb{E}[T_\ell^k(s_j^\ell,a')\mathbb{I}\{\xi_{Z,K}^C\}].$$

Step 3 (Sampling of oracle for an episode k when $\widehat{Z}^k \geq 0$): First note that when $\widehat{Z}^k \geq 0$ the oracle samples at episode k according to the policy $\mathbf{b}_*$. The following the same steps as in step 3 of Theorem 2 we can show that. At episode k, time $\ell+1$, the $\mathbf{b}_*$ samples the state-action tuple, action $\arg\max_a U_{\ell+1}^k(s_i^{\ell+1},a)$ where

$$U_\ell^k(s_i^\ell,a) \coloneqq \frac{\mathbf{b}_{*,\ell}(a|s_i^\ell)}{T_\ell^k(s_i^\ell,a)} \tag{41}$$

Let $\ell+1 > 2SA$ be the time at which a given state-action (s_i^ℓ,p') is visited for the last time, i.e., $T_\ell^k(p') = T_L^K(p') - 1$ and $T_{\ell+1}^k(p') = T_L^K(p')$. Note that as $n = KL \geq 4SA$, there is at least one state-action pair (s_i^ℓ,p') such that this happens, i.e. such that it is visited after the initialization phase. Since the oracle chooses to pull visit (s_i^ℓ,p') at time $\ell+1$, we have for any state-action pair (s_i^ℓ,p')

$$U_{\ell+1}^k(s_i^{\ell+1},p) \leq U_{\ell+1}^k(s_i^{\ell+1},p'). \tag{42}$$

From (41) and using the fact that $T_\ell^k(s_i^\ell,p') = T_L^K(s_i^\ell,p') - 1$, we can show that

$$U_{\ell+1}^k(s_i^{\ell+1},p') \leq \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_t^k(s_i^{\ell+1},p')} = \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_L^K(s_i^{\ell+1},p') - 1} \tag{43}$$

Also note that

$$U_{\ell+1}^k(s_i^{\ell+1},p) = \frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_t^k(s_i^{\ell+1},p)} \overset{(a)}{\geq} \frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_L^K(s_i^{\ell+1},p)}. \tag{44}$$

where, (a) follows as $T_t(p) \leq T_L^K(p, s_i^{\ell+1})$ (i.e., the number of times p has been sampled can only increase after time ℓ). Combining (42), (43), (44) we can show that for any action p:

$$\frac{\mathbf{b}_*(p|s_i^{\ell+1})}{T_L^K(p, s_i^{\ell+1})} \leq \frac{\mathbf{b}_*(p'|s_i^{\ell+1})}{T_L^K(p', s_i^{\ell+1}) - 1} \tag{45}$$

Note that in the above equation, there is no dependency on ℓ, and thus, the probability that (45) holds for any $(s_i^{\ell+1}, p)$ and for any $(s_i^{\ell+1}, p')$ such that state-action $(s_i^{\ell+1}, p')$ is visited after the initialization phase, i.e., such that $T_L^K(s_i^{\ell+1}, p') > 2$ depends on the probability of event $\xi_{Z,n}$.

Step 4. (Lower bound on $T_L^K(s_i^\ell, p)$ for $\widehat{Z}^k \geq 0$): If a state-action tuple s_i^ℓ, p, p is under-pulled compared to its optimal allocation without taking into account the initialization phase, i.e., $T_L^K(s_i^\ell, p) - 2 < \mathbf{b}(p|s_i^\ell)(n-2A)$, then from the constraint $\sum_{p'} \left(T_L^K(s, p') - 2\right) = n - 2SA$ and the definition of the optimal allocation, we deduce that there exists at least another state-action tuple s_i^ℓ, p' that is over-visited compared to its optimal allocation without taking into account the initialization phase, i.e., $T_L^K(s_i^\ell, p') - 2 > \mathbf{b}(s_i^\ell, p')(n - 2SA)$. Note that for this action, $T_L^K(s_i^\ell, p') - 2 > \mathbf{b}_*(p'|s_i^\ell)(n - 2SA) \geq 0$, so we know that this specific action is pulled at least once after the initialization phase and that it satisfies (45). Recall that we have defined $M(s_i^\ell) = \sum_a \pi(a|s_i^\ell)\sigma(s_i^\ell, a)$. Further define $M = \sum_{\ell=1}^L \sum_{s_i^\ell} M(s_i^\ell)$. Using the definition of the optimal allocation $T_L^{*,K}(s_i^\ell, p') = n_f \frac{\mathbf{b}_*(p'|s_i^\ell)}{M(s_i^\ell)}$, and the fact that $T_L^K(s_i^\ell, p') \geq \mathbf{b}_*(p'|s_i^\ell)(n_f - 2SA) + 2$, (45) may be written as for any state-action tuple (s_i^ℓ, p)

$$\frac{\mathbf{b}_*(p|s_i^\ell)}{T_L^K(s_i^\ell, p)} \leq \frac{\mathbf{b}_*(p'|s_i^\ell)}{T_L^{*,K}(p', s_i^\ell)} \frac{n_f}{(n_f - 2SA)} \leq \frac{M(s_i^\ell)}{n_f} + \frac{4AM(s_i^\ell)}{n_f^2} \tag{46}$$

because $n_f \geq 4SA$. By rearranging (46), we obtain the lower bound on $T_L^K(s_i^\ell, p)$:

$$T_L^K(s_i^\ell, p) \geq \frac{\mathbf{b}_*(p|s_i^\ell)}{\frac{M(s_i^\ell)}{n_f} + \frac{4AM(s_i^\ell)}{n_f^2}} = \frac{\mathbf{b}_*(p|s_i^\ell)}{\frac{M(s_i^\ell)}{n_f}\left(1 + \frac{4A}{n_f}\right)} \overset{(a)}{\geq} T_L^{*,K}(s_i^\ell, p) - 4A\mathbf{b}_*(p|s_i^\ell), \tag{47}$$

where in (a) we use $1/(1+x) \geq 1 - x$ (for $x > -1$). Note that the lower bound holds on $\xi_{c,K}$ for any action p.

Step 5. (Upper bound on $T_L^K(s_i^\ell, p)$ for $\widehat{Z}^k \geq 0$): Now using (47) and the fact that n_f is given by $\sum_{\ell=1}^L \sum_{s_j^\ell} \sum_{a'=1}^A \mathbb{E}[T_L^K(s_j^\ell, a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] = n_f$, we obtain

$$T_L^K(s_i^\ell, p) = n_f - \sum_{p' \neq p} T_L^K(s_i^\ell, p') \leq \left(n_f - \sum_{p' \neq p} T_L^{*,K}(s_i^\ell, p')\right) + \sum_{p' \neq p} 4A\mathbf{b}_*(p'|s_i^\ell).$$

Now since $\sum_{p' \neq p} \mathbf{b}_*(p'|s_i^\ell) \leq 1$ we can show that

$$T_L^K(s_i^\ell, p) \leq T_L^{*,K}(s_i^\ell, p) + 4A. \tag{48}$$

Step 6 (Bound part A): We now bound the part A using (46)

$$\begin{aligned}
&\sum_a \pi^2(a|s_1^1)\left[\frac{\sigma^2(s_1^1, a)}{\underline{T}_L^{(2),K}(s_1^1, a)}\right]\mathbb{E}[T_L^K(s_1^1, a)\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
&\quad + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^L \sum_{s_j^\ell} P(s_j^\ell|s_1^1, a) \sum_{a'} \pi^2(a'|s_j^\ell)\left[\frac{\sigma^2(s_j^\ell, a')}{\underline{T}_L^{(2),K}(s_j^\ell, a')}\right]\mathbb{E}[T_L^K(s_j^\ell, a')\mathbb{I}\{\xi_{Z,K}\} \cap \mathbb{I}\{\xi_{c,K}\}] \\
&\overset{(a)}{\leq} \left(\frac{M(s_1^1)}{n_f} + \frac{4AM(s_1^1)}{n_f^2}\right)^2 n_f + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^L \sum_{s_j^\ell} P(s_j^\ell|s_1^1, a)\left(\frac{M(s_j^\ell)}{n_f} + \frac{4AM(s_j^\ell)}{n_f^2}\right)^2 n_f \\
&= \frac{M^2(s_1^1)}{n_f} + \frac{8AM^2(s_1^1)}{n_f^2} + \frac{16A^2M^2(s_1^1)}{n_f^3} \\
&\quad + \gamma^2 \sum_a \pi^2(a|s_1^1) \sum_{\ell=2}^L \sum_{s_j^\ell} P(s_j^\ell|s_1^1, a)\left(\frac{M^2(s_j^\ell)}{n_f} + \frac{8AM^2(s_j^\ell)}{n_f^2} + \frac{16A^2M^2(s_j^\ell)}{n_f^3}\right)
\end{aligned}$$

where, in (a) follows from the definition of $M(s)$ and n_f.

Step 7 (Upper bound to Constraint violation): In this step we bound the quantity $\mathcal{C}_n^*(\pi) = \sum_{j=1}^k \mathbb{I}\{\widehat{Z}^j < 0, \mathbf{b}^j \in \{\mathbf{b}_*, \pi_0\}\}$. Define the number of times the policy $\mathbf{b}_*$ is played till episode k is $T^k(\mathbf{b}_*)$ and the number of times the baseline policy is played is given by $T^k(\pi_0)$. Observe that $\mathcal{C}_n^*(\pi) = \sum_{j=1}^k \mathbb{I}\{\widehat{Z}^j < 0, \mathbf{b}^j \in \{\mathbf{b}_*, \pi_0\}\} = T^K(\pi_0)\mathbb{I}\{\xi_{Z,K}^C\}$ as when the constraint are violated and policy π_0 is played. Let $\tau = \max\left\{k \le K \text{ and } n_f \ge \frac{\log(SAn(n+1)/\delta)}{\min_{s,a} \mathbf{b}_*(a|s)\Delta^{c,\alpha,(2)}(s,a)} \mid \mathbf{b}^k = \pi_0\right\}$ be the last episode in which the baseline policy is played. We will define formally the gap $\Delta^{c,\alpha,(2)}(s,a)$ later. Observe that the constraint violation can be re-stated as follows:

$$
\begin{aligned}
&\sum_{k=1}^{\tau} Y_{\mathbf{b}^k}^c(s_1^1) \coloneqq \sum_{k=1}^{\tau}\sum_a \mathbf{b}^k(a|s_1^1)\left(\widehat{\mu}_L^{c,k}(s_1,a) + \sum_{s_j^2} P(s_j^2|s_1^1,a) Y_{\mathbf{b}^k}^c(s_j^2)\right) < (1-\alpha)\tau V_{\pi_0}^c(s_1^1)\\
\implies &\sum_{k=1}^{\tau}\sum_a \mathbf{b}^k(a|s_1^1)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a) + \sum_{s_j^2} P(s_j^2|s_1^1,a) \underline{Y}_{\mathbf{b}^k}^c(s_j^2)\right) < (1-\alpha)\tau V_{\pi_0}^c(s_1^1)\\
\overset{(a)}{\implies} &\sum_{k=1}^{\tau}\sum_a \mathbf{b}^k(a|s_1^1)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a) + \sum_{s_j^2} P(s_j^2|s_1^1,a) \underline{Y}_{\mathbf{b}^k}^c(s_j^2)\right)\\
&\qquad < (1-\alpha)\sum_{k=1}^{\tau}\pi_0(0|s_1^1)\left(\mu^c(s_1^1,0) + \sum_{s_j^2} P(s_j^2|s_1^1,0) V_{\pi_0}^c(s_j^2)\right)\\
\implies &\sum_{k=1}^{\tau}\sum_a T_L^k(s_1^1,a)\left(\underline{\widehat{\mu}}_L^{c,k}(s_1^1,a) + \sum_{s_j^2} P(s_j^2|s_1^1,a) \underline{Y}_{\mathbf{b}^k}^c(s_j^2)\right)\\
&\qquad < (1-\alpha)\sum_{k=1}^{\tau} T_L^k(s_1^1,a)\left(\mu^c(s_1^1,0) + \sum_{s_j^2} P(s_j^2|s_1^1,0) V_{\pi_0}^c(s_j^2)\right)\\
\overset{(b)}{\implies} &\underbrace{\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a)}_{\textbf{Part A}} + \sum_a T_L^\tau(s_1^1,a)\sum_{s_j^2} P(s_j^2|s_1^1,a)\underline{Y}_{\mathbf{b}^k}^c(s_j^2)\\
&\qquad < \underbrace{(1-\alpha)\sum_a T_L^\tau(s_1^1,0)\mu^c(s_1^1,0)}_{\textbf{Part B}} + (1-\alpha)T_L^\tau(s_1^1,0)\sum_{s_j^2} P(s_j^2|s_1^1,0) V_{\pi_0}^c(s_j^2)
\end{aligned}
\tag{49}
$$

where (a) follows as π_0 samples baseline action 0 for each state $s \in [S]$, and in (b) the $T_L^\tau(s_1^1,a)$ denotes the total samples of state-action tuple till episode τ. Comparing **Part A** and **Part B** for level $\ell = 1$ we observe that the constraint violation must satisfy

$$\sum_a T_L^\tau(s_1^1,a)\underline{\widehat{\mu}}_L^{c,\tau}(s_1^1,a) < (1-\alpha)T_L^\tau(s_1^1,0)\mu^c(s_1^1,0)$$

which can be reduced by following the same way as step 7 as Theorem 2

$$T_L^{\tau-1}(s_1^1,0) \le \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1 + \sum_{a=1}^A N(s_1^1,a)\right).$$

where $\Delta^{c,\alpha}(s_1^1,a) \coloneqq (1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a)$ and

$$
\begin{aligned}
N(s_1^1,a) &\coloneqq T_L^{\tau-1}(s_1^1,a)\cdot\left((1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)/T_L^{\tau-1}(s_1^1,a)}\right)\\
&= \Delta^{c,\alpha}(s_1^1,a)T_L^{\tau-1}(s_1^1,a) + c_1\sqrt{\log(An(n+1)/\delta)T_L^{\tau-1}(s_1^1,a)}
\end{aligned}
\tag{50}
$$

is a bound on the decrease in $\widehat{Z}_\tau$ in the first $\tau-1$ rounds due to choosing action a in s_1^1. We will now bound $N(s_1^1,a)$ for each a. Now observe

$$\begin{aligned}\Delta^{c,\alpha}(s_1^1,a) = (1-\alpha)\mu^c(s_1^1,0) - \mu^c(s_1^1,a) &= \mu^c(s_1^1,0) - \alpha\mu^c(s_1^1,0) - \mu^c(s_1^1,a)\\ &= -(\mu^{*,c}(s_1^1) - \mu^c(s_1^1,0)) - \alpha\mu^c(s_1^1,0) + (\mu^{*,c}(s_1^1) - \mu^c(s_1^1,a))\\ &= -\Delta^c(s_1^1,0) - \alpha\mu^c(s_1^1,0) + \Delta^c(s_1^1,a).\end{aligned}$$

where, $\mu^{*,c}(s_1^1) = \max_a \mu^c(s_1^1,a)$. It follows then that using step 7 as Theorem 2 for the state s_1^1

$$n_u(s_1^1) \le \frac{1}{\alpha\mu^c(s_1^1,0)}\left(1+\sum_{a=1}^{A} N(s_1^1,a)\right) \le \frac{H_{*,(2)}(s_1^1)}{2}\frac{n}{M(s_1^1)}$$

where

$$\begin{aligned}H_{*,(2)}(s_i^\ell) &\coloneqq \sum_a \mathbf{b}_*(a|s_i^\ell)\overset{+}{\min}\{\Delta^c(s_i^\ell,a), \Delta^c(s_i^\ell,0)-\Delta^c(s_i^\ell,a)\},\\ M(s_i^\ell) &\coloneqq \sum_a \sqrt{\pi^2(a|s_i^\ell)\left(\sigma^2(s_i^\ell,a)+\sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell,a)M^2(s_j^{\ell+1})\right)}\end{aligned} \tag{51}$$

Similarly, for an arbitrary level $\ell \in [L]$, we can show using (49) that the constraint violation must satisfy

$$\begin{aligned}&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a T_L^\tau(s_i^{\ell'},a)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}} T_L^\tau(s_i^{\ell'},0)\mu^c(s_i^{\ell'},0)\\
\overset{(a)}{\Longrightarrow}&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a \left(T_L^{*,K}(s_i^{\ell'},a) - 4A\mathbf{b}_*(a|s_i^{\ell'})\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)+4A\right)\mu^c(s_i^{\ell'},0)\\
\Longrightarrow&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a \left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a)\\
&< (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right)\mu^c(s_i^{\ell'},0) + 8LSA^2(\mu^c(s_i^{\ell'},0)+\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a))\\
\Longrightarrow&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a \left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a)\\
&< (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right)\mu^c(s_i^{\ell'},0) + 8LSA^2(\mu^c(s_i^{\ell'},0)+\widehat{\mu}_L^{c,\tau}(s_i^{\ell'},a) - \sqrt{\frac{\log((SAn(n+1)/\delta)}{2T_L^\tau(s_i^{\ell'},a)}})\\
\overset{(b)}{\Longrightarrow}&\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a \left(T_L^{*,K}(s_i^{\ell'},a)\right)\underline{\widehat{\mu}}_L^{c,\tau}(s_i^{\ell'},a) < (1-\alpha)\max_{s,a}\mu_0^c(s,a)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'},0)\right) + 16LSA^2\end{aligned} \tag{52}$$

where, (a) follows from (48) and (b) follows as $\mu(s,a) \in (0,1]$ for all s,a. It follows then that using step 7 of Theorem 2 and definition of $N(s_j^\ell)$ from (50)

$$\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}} T_L^{*,K}(s_i^{\ell'},0) \le \frac{1}{\alpha\max_s\mu^c(s,0)}\left(1+\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a N(s_j^\ell,a)\right) \le \frac{n}{2}\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a \frac{H_{*,(2)}(s_i^{\ell'})}{M(s_i^{\ell'})} + 16LSA^2$$

which gives a bound on how many times action $\{0\}$ is sampled across different states till level ℓ. Summing over all states s_j^ℓ till level L we can show that

$$n_u = \sum_{\ell=1}^{L}\sum_{s_j^\ell} T_L^{*,K}(s_j^\ell, 0) \leq \frac{n}{2}\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)} + 16LSA^2 \overset{(a)}{\leq} \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2 \tag{53}$$

where, in (a) we define $M_{\min} = \min_s M(s)$, and $H_{*,(2)} = \sum_{\ell=1}^{L}\sum_{s_j^\ell} H_{*,(2)}(s_j^\ell)$. Finally, observe that $16LSA^2$ does not depend on the episode K.

Step 8 (Lower bound to Constraint violation): For the lower bound to the constraint we equate Equation (49) to 0 and show that

$$\underbrace{\sum_a T_L^\tau(s_1^1, a)\widehat{\underline{\mu}}_L^{c,\tau}(s_1^1, a)}_{\textbf{Part A}} + \sum_a T_L^\tau(s_1^1, a)\sum_{s_j^2} P(s_j^2|s_1^1, a)\underline{V}_{\mathbf{b}^k}^c(s_j^2)$$
$$= \underbrace{(1-\alpha)\sum_a T_L^\tau(s_1^1, 0)\mu^c(s_1^1, 0)}_{\textbf{Part B}} + (1-\alpha)T_L^\tau(s_1^1, 0)\sum_{s_j^2} P(s_j^2|s_1^1, 0)V_{\pi_0}^c(s_j^2)$$

Again comparing **Part A** and **Part B** for level $\ell = 1$ we observe that the lower bound to constraint violation must satisfy

$$\sum_a T_L^\tau(s_1^1, a)\widehat{\underline{\mu}}_L^{c,\tau}(s_1^1, a) = (1-\alpha)T_L^\tau(s_1^1, 0)\mu^{c,}(s_1^1, 0)$$

which can be reduced by following the same way as step 8 as Theorem 2

$$\sum_a T_L^{\tau-1}(s_1^1, 0) \geq \frac{1}{\alpha\mu^c(s_1^1, 0)}\left(1 + \sum_{a=1}^{A}\underline{N}(s_1^1, a)\right).$$

where $\Delta^{c,\alpha}(s_1^1, a) \coloneqq (1-\alpha)\mu^c(s_1^1, 0) - \mu^c(s_1^1, a)$ and

$$\begin{aligned}\underline{N}(s_1^1, a) &\coloneqq T_L^{\tau-1}(s_1^1, a)\cdot\left((1-\alpha)\mu^c(s_1^1, 0) - \mu^c(s_1^1, a) + c_1\sqrt{\log(An(n+1)/\delta)/T_L^{\tau-1}(s_1^1, a)}\right)\\ &= \Delta^{c,\alpha}(s_1^1, a)T_L^{\tau-1}(s_1^1, a) + c_1\sqrt{\log(An(n+1)/\delta)T_L^{\tau-1}(s_1^1, a)}\\ &\overset{(a)}{\geq} \Delta^{c,\alpha}(s_1^1, a)\left(T_L^{*,K}(s_1^1, a) - 4A\mathbf{b}_*(a|s_1^1)\right) + c_1\sqrt{\log(An(n+1)/\delta)\left(T_L^{*,K}(s_1^1, a) - 4A\mathbf{b}_*(a|s_1^1)\right)}\end{aligned}$$

where, (a) follows from (47). Then following the same way as step 8 of Theorem 2 we can show that

$$T_L^{\tau-1}(s_1^1, 0) \geq \frac{1}{\alpha\mu^c(s_1^1, 0)}\left(1 + \sum_{a=1}^{A}\underline{N}(s_1^1, a)\right) \geq \frac{n_f}{M(s_1^1)}\left(\frac{H_{*,(2)}(s_1^1)}{8} - \frac{A}{2}\frac{H_{*,(2)}(s_1^1)}{M(s_1^1)}\right) - 16SA$$

Similarly for any arbitrary level $\ell \in [L]$ following the same way as step 7 above it can be shown that

$$\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a\left(T_L^{*,K}(s_i^{\ell'}, a) + 4A\right)\widehat{\underline{\mu}}_L^{c,\tau}(s_i^{\ell'}, a) \geq (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\left(T_L^{*,K}(s_i^{\ell'}, 0) - 4A\mathbf{b}_*(0|s_i^{\ell'})\right)\mu^c(s_i^{\ell'}, 0)$$
$$\implies \sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}}\sum_a T_L^{*,K}(s_j^\ell, a)\widehat{\underline{\mu}}_L^{c,\tau}(s_j^\ell, a) \geq (1-\alpha)\sum_{\ell'=1}^{\ell}\sum_{s_i^{\ell'}} T_L^{*,K}(s_i^\ell, 0)\mu^c(s_i^\ell, 0) - 16LSA^2$$

Again following the same way as step 8 of Theorem 2 for the state s_j^ℓ, the lower bound to the total number of times the

baseline actions are sampled across states till level ℓ is given by we can show that

$$\sum_{\ell'=1}^{\ell}\sum_{s_j^{\ell'}} T_L^{*,K}(s_j^{\ell'},0) \geq \frac{1}{\alpha \max_{s_j^\ell}\mu^c(s_j^\ell,0)}\left(1+\sum_{\ell'=1}^{\ell}\sum_{s_j^{\ell'}}\sum_{a=1}^{A}\underline{N}(s_j^{\ell'},a)\right)$$
$$\geq \sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}-\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)-16LSA^2$$

Finally summing over all states s_j^ℓ and level L we can show that

$$\sum_{\ell=1}^{L}\sum_{s_j^\ell} T_L^{*,K}(s_j^\ell,0) \geq \sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}-\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)-16LSA^2 \tag{54}$$

Again, observe that $16LSA^2$ does not depend on the episode K.

Step 9 (Bound Part B): Then from (54) we can show that

$$\frac{M(s_1^1)}{\sum_{\ell=1}^{L}\sum_{s_j^\ell}T_L^{*,K}(s_j^\ell,0)} \leq \frac{M(s_1^1)}{\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{n_f}{M(s_j^\ell)}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}-\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)-16LSA^2}$$
$$\overset{(a)}{\leq} (M(s_1^1)+16LSA^2)\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{M(s_j^\ell)}{n_f}\left(\frac{H_{*,(2)}(s_j^\ell)}{8}+\frac{A}{2}\frac{H_{*,(2)}(s_j^\ell)}{M(s_j^\ell)}\right)$$
$$\leq (M(s_1^1)+16LSA^2)\sum_{\ell=1}^{L}\sum_{s_j^\ell}\frac{M(s_j^\ell)}{n_f}\left(2+H_{*,(2)}(s_j^\ell)\right)$$
$$\overset{(b)}{\leq} (M(s_1^1)+16LSA^2)\frac{M}{n_f}\left(2+H_{*,(2)}\right)$$

where, (a) follows for $1/(x-c)\leq x+c$ for $x^2\geq 1+c^2$ and $c>0$. The (b) follows for $M=\sum_{\ell=1}^{L}\sum_{s_j^\ell}M(s_j^\ell)$, and $H_{*,(2)}=\sum_{\ell=1}^{L}\sum_{s_j^\ell}H_{*,(2)}(s_j^\ell)$. It follows then by setting $n_f=n-n_u$ that

$$\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{Z,K}^C\}\right] \overset{(a)}{\leq} \left(\frac{(M(s_1^1)+16LSA^2)M\left(2+H_{*,(2)}\right)}{n_f}\right)^2 n_u$$
$$\overset{(b)}{=} \frac{(M(s_1^1)+16LSA^2)^2 n_u}{(n-n_u)^2}\left(2+H_{*,(2)}\right)^2$$
$$\overset{(c)}{\leq} \frac{(M(s_1^1)+16LSA^2)^2 H_{*,(2)}n}{(n-H_{*,(2)}n)^2}\left(2+H_{*,(2)}\right)^2$$
$$\leq \frac{M^2(s_1^1)}{n}\left(32MLSA^2+H_{*,(2)}\right)^2$$

where, (a) follows from Lemma A.1, (b) follows from the definition of $H_{*,(2)}$, and (c) follows from (53).

Step 10 (Combine everything): Combining everything from step 5, step 8 and setting $\delta=1/n^2$ we can show that the MSE of oracle scales as

$$\mathcal{L}_n(\pi_{\mathbf{b}}^k,\mathbf{b}_*^k) \leq \frac{M^2(s_1^1)}{n}+\frac{8AM^2(s_1^1)}{n^2}+\frac{16A^2M^2(s_1^1)}{n^3}+\frac{M^2(s_1^1)}{n}\left(32MLSA^2+H_{*,(2)}\right)^2+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1)-V_\pi(s_1^1)\right)^2\mathbb{I}\{\xi_{c,K}^C\}\right]}_{\textbf{Part C, Safety event does not hold}}$$
$$\overset{(a)}{\leq} \frac{M^2(s_1^1)}{n}+\frac{8AM^2(s_1^1)}{n^2}+\frac{16A^2M^2(s_1^1)}{n^3}+\frac{M^2(s_1^1)}{n}\left(32MLSA^2+H_{*,(2)}\right)^2+2\sum_{t=1}^{n}\frac{2\eta+4\eta^2}{n^2} \tag{55}$$

where, (a) follows as $\mathbb{E}_{\mathcal{D}}\left[\left(Y_n(s_1^1) - V_\pi(s_1^1)\right)^2 \mathbb{I}\{\xi_{c,K}^C\}\right] \leq 2\eta + 4\eta^2$ and using the low error probability of the constraint event from Lemma F.4. The claim of the proposition follows. □

E.1. Tree Regret Corollary

Corollary 1. *Under Assumption 3.2 the constraint regret in the Tree MDP is given by* $\overline{\mathcal{R}}_n^c \leq O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right)$ *and the regret is given by* $\overline{\mathcal{R}}_n \leq O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right)$.

Proof. The upper bound to the safe oracle constraint is given by (53) as follows

$$\mathcal{C}_n^*(\pi, \mathbf{b}_*^k) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2.$$

The upper bound to the constraint violation of SaVeR is given by (36)

$$\mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16LSA^2 + O\left(\frac{(2\eta + 4\eta^2)L^2S^2A^4H_{*,(2)}^2M^2\sqrt{\log(SAn(n+1)/\delta)}}{\min_s \mathbf{b}^{*,k,(3/2)}(s) n^{3/2}}\right).$$

Hence, from the constraint regret definition, we can show that

$$\overline{\mathcal{R}}_n^c = \mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{C}}_n^*(\pi, \mathbf{b}_*) \leq O\left(\frac{\log n}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right).$$

Observe that the loss of the agnostic algorithm SaVeR is given by (38) and the upper bound to the oracle loss is given by (55). Comparing these two losses directly leads to the regret as follows:

$$\overline{\mathcal{R}}_n = \mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{L}}_n^*(\pi, \mathbf{b}_*^k) = O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right).$$

The claim of the corollary follows. □

Corollary 2. *Under Assumption 3.2 the constraint regret in the bandit setting is given by* $\overline{\mathcal{R}}_n^c \leq O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right)$ *and the regret is given by* $\overline{\mathcal{R}}_n \leq O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right)$.

Proof. The bandit setting consists of a single state, and so we can define the quantity $H_{*,(2)} = \frac{1}{\alpha\mu(0)}\sum_{a\in\mathcal{A}\setminus\{0\}} \pi(a)\sigma(a)\min^+\{\Delta^c(a), \Delta^c(0) - \Delta^c(a)\}$ The upper bound to the oracle constraint is given by (53) as follows

$$\mathcal{C}_n^*(\pi, \mathbf{b}_*^k) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16A^2.$$

The upper bound to the constraint violation of SaVeR is given by (36)

$$\mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) \leq \frac{H_{*,(2)}}{2}\frac{n}{M_{\min}} + 16A^2 + O\left(\frac{(2\eta + 4\eta^2)A^4H_{*,(2)}^2M^2\sqrt{\log(An(n+1)/\delta)}}{\min_s \mathbf{b}^{*,k,(3/2)}(s) n^{3/2}}\right).$$

Hence, from the constraint regret definition, we can show that

$$\overline{\mathcal{R}}_n^c = \mathcal{C}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{C}}_n^*(\pi, \mathbf{b}_*^k) \leq O\left(\frac{\log n}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right).$$

Observe that the loss of the agnostic algorithm SaVeR is given by (38) and the upper bound to the oracle loss is given by (55). Comparing these two losses directly leads to the regret as follows:

$$\overline{\mathcal{R}}_n = \mathcal{L}_n(\pi, \widehat{\mathbf{b}}^k) - \overline{\mathcal{L}}_n^*(\pi, \mathbf{b}_*^k) = O\left(\frac{\log(n)}{\mathbf{b}_{*,\min}^{3/2} n^{3/2}}\right).$$

The claim of the corollary follows. □

F. Support Lemmas

Lemma F.1. ***(Hoeffding's Lemma)****(Massart, 2007) Let Y be a real-valued random variable with expected value $\mathbb{E}[Y] = \mu$, such that $a \leq Y \leq b$ with probability one. Then, for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq \exp\left(\lambda\mu + \frac{\lambda^2(b-a)^2}{8}\right)$$

Lemma F.2. ***(Concentration lemma 1)*** *Let $V_t = R_t(s,a) - \mathbb{E}[R_t(s,a)]$ and be bounded such that $V_t \in [-\eta, \eta]$. Let the total number of times the state-action (s,a) is sampled be T. Then we can show that for an $\epsilon > 0$*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} R_t(s,a) - \mathbb{E}[R_t(s,a)]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2 T}{\eta^2}\right).$$

Proof. Let $V_t = R_t(s,a) - \mathbb{E}[R_t(s,a)]$. Note that $\mathbb{E}[V_t] = 0$. Hence, for the bounded random variable $V_t \in [-\eta, \eta]$ we can show from Hoeffding's lemma in Lemma F.1 that

$$\mathbb{E}[\exp(\lambda V_t)] \leq \exp\left(\frac{\lambda^2}{8}(\eta - (-\eta))^2\right) \leq \exp\left(2\lambda^4\eta^2\right)$$

Let s_{t-1} denote the last time the state s is visited and action a is sampled. Observe that the reward $R_t(s,a)$ is conditionally independent and η^2-sub-Gaussian. Next we can bound the probability of deviation as follows:

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{T}(R_t(s,a) - \mathbb{E}[R_t(s,a)]) \geq \epsilon\right) &= \mathbb{P}\left(\sum_{t=1}^{T} V_t \geq \epsilon\right) \\
&\overset{(a)}{=} \mathbb{P}\left(e^{\lambda\sum_{t=1}^{T} V_t} \geq e^{\lambda\epsilon}\right) \\
&\overset{(b)}{\leq} e^{-\lambda\epsilon}\mathbb{E}\left[e^{-\lambda\sum_{t=1}^{T} V_t}\right] \\
&= e^{-\lambda\epsilon}\mathbb{E}\left[\mathbb{E}\left[e^{-\lambda\sum_{t=1}^{T} V_t}\middle| s_{T-1}\right]\right] \\
&\overset{(c)}{=} e^{-\lambda\epsilon}\mathbb{E}\left[\mathbb{E}\left[e^{-\lambda V_T}|S_{T-1}\right]\mathbb{E}\left[e^{-\lambda\sum_{t=1}^{T-1} V_t}\middle| s_{T-1}\right]\right] \\
&\leq e^{-\lambda\epsilon}\mathbb{E}\left[\exp\left(2\lambda^4\eta^2\right)\mathbb{E}\left[e^{-\lambda\sum_{t=1}^{T-1} V_t}\middle| s_{T-1}\right]\right] \\
&= e^{-\lambda\epsilon}e^{2\lambda^2\eta^2}\mathbb{E}\left[e^{-\lambda\sum_{t=1}^{T-1} V_t}\right] \\
&\vdots \\
&\overset{(d)}{\leq} e^{-\lambda\epsilon}e^{2\lambda^2 T\eta^2} \\
&\overset{(e)}{\leq} \exp\left(-\frac{2\epsilon^2}{T\eta^2}\right)
\end{aligned} \tag{56}$$

where (a) follows by introducing $\lambda \in \mathbb{R}$ and exponentiating both sides, (b) follows by Markov's inequality, (c) follows as V_t is conditionally independent given s_{T-1}, (d) follows by unpacking the term for T times and (e) follows by taking

$\lambda = \epsilon/4T\eta^2$. Hence, it follows that

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} R_t(s,a) - \mathbb{E}[R_t(s,a)]\right| \geq \epsilon\right) = \mathbb{P}\left(\sum_{t=1}^{T}(R_t(s,a) - \mathbb{E}[R_t(s,a)]) \geq T\epsilon\right) \overset{(a)}{\leq} 2\exp\left(-\frac{2\epsilon^2 T}{\eta^2}\right).$$

where, (a) follows by (56) by replacing ϵ with ϵT, and accounting for deviations in either direction. □

Lemma F.3. ***(Concentration lemma 2)*** *Let $\mu^2(s,a) = \mathbb{E}\left[R_t^2(s,a)\right]$. Let $R_t(s,a)$ be η^2 sub-Gaussian. Let $n = KL$ be the total budget of state-action samples. Define the event*

$$\xi_\delta = \left(\bigcap_{s\in\mathcal{S}}\bigcap_{1\leq a\leq A, T_n(s,a)\geq 1}\left\{\left|\frac{1}{T_n(s,a)}\sum_{t=1}^{T_n(s,a)} R_t^2(s,a) - \mu^2(s,a)\right| \leq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}}\right\}\right) \bigcap$$
$$\left(\bigcap_{s\in\mathcal{S}}\bigcap_{1\leq a\leq A, T_n(s,a)\geq 1}\left\{\left|\frac{1}{T_n(s,a)}\sum_{t=1}^{T_n(s,a)} R_t(s,a) - \mu(s,a)\right| \leq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}}\right\}\right) \quad (57)$$

Then we can show that $\mathbb{P}(\xi_\delta) \geq 1 - 2\delta$.

Proof. First note that the total budget $n = KL$. Observe that the random variable $R_t^k(s,a)$ and $R_t^{(2),k}(s,a)$ are conditionally independent given the previous state S_{t-1}^k. Also observe that for any $\eta > 0$ we have that $R_t^k(s,a), R_t^{(2),k}(s,a) \leq 2\eta + 4\eta^2$, where $R_t^{(2),k}(s,a) = (R_t^k(s,a))^2$. Hence we can show that

$$\mathbb{P}\left(\bigcap_{s\in\mathcal{S}}\bigcap_{1\leq a\leq A, T_n(s,a)\geq 1}\left\{\left|\frac{1}{T_n(s,a)}\sum_{t=1}^{T_n(s,a)} R_t^2(s,a) - \mu^2(s,a)\right| \geq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}}\right\}\right)$$
$$\leq \mathbb{P}\left(\bigcup_{s\in\mathcal{S}}\bigcup_{1\leq a\leq A, T_n(s,a)\geq 1}\left\{\left|\frac{1}{T_n(s,a)}\sum_{t=1}^{T_n(s,a)} R_t^2(s,a) - \mu^2(s,a)\right| \geq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}}\right\}\right)$$
$$\overset{(a)}{\leq} \sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{t=1}^{n}\sum_{T_n(s,a)=1}^{t} 2\exp\left(-\frac{2T_n}{4(\eta^2+\eta)^2}\cdot\frac{4(\eta^2+\eta)^2\log(SAn(n+1)/\delta)}{2T_n(s,a)}\right) = \delta.$$

where, (a) follows from Lemma F.2. Note that in (a) we have to take a double union bound summing up over all possible pulls T_n from 1 to n as T_n is a random variable. Similarly we can show that

$$\mathbb{P}\left(\bigcap_{s\in\mathcal{S}}\bigcap_{1\leq a\leq A, T_n(s,a)\geq 1}\left\{\left|\frac{1}{T_n(s,a)}\sum_{t=1}^{T_n(s,a)} R_t(s,a) - \mu(s,a)\right| \geq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}}\right\}\right)$$
$$\overset{(a)}{\leq} \sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{t=1}^{n}\sum_{T_n(s,a)=1}^{t} 2\exp\left(-\frac{2T_n}{4(\eta^2+\eta)^2}\cdot\frac{4(\eta^2+\eta)^2\log(SAn(n+1)/\delta)}{2T_n(s,a)}\right) = \delta.$$

where, (a) follows from Lemma F.2. Hence, combining the two events above we have the following bound

$$\mathbb{P}(\xi_\delta) \geq 1 - 2\delta.$$

□

Corollary 3. *Under the event ξ_δ in (57) we have for any state-action pair in an episode k the following relation with probability greater than $1-\delta$*

$$|\widehat{\sigma}_t^k(s,a) - \sigma(s,a)| \leq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_L^K(s,a)}}.$$

where, $T_L^K(s,a)$ is the total number of samples of the state-action pair (s,a) till episode k.

Proof. Observe that the event ξ_δ bounds the sum of rewards $R_t^k(s,a)$ and squared rewards $R_t^{k,(2)}(s,a)$ for any $T_L^K(s,a) \geq 1$. Hence we can directly apply the Lemma F.3 to get the bound. □

Lemma F.4. *Let $\mu^c(s,a) = \mathbb{E}\left[C_t(s,a)\right]$ and $C_t(s,a) \leq 2\eta$. Define the event*

$$\overline{\xi}_\delta = \bigcap_{s\in\mathcal{S}} \bigcap_{1\leq a\leq A, T_n(s,a)\geq 1} \left\{ \left| \frac{1}{T_n(s,a)} \sum_{t=1}^{T_n(s,a)} C_t(s,a) - \mu^c(s,a) \right| \leq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}} \right\}. \tag{58}$$

Then we can show that $\mathbb{P}(\overline{\xi}_\delta) \geq 1-\delta$.

Proof. We can show that

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{s\in\mathcal{S}} \bigcap_{1\leq a\leq A, T_n(s,a)\geq 1} \left\{ \left| \frac{1}{T_n(s,a)} \sum_{t=1}^{T_n(s,a)} C_t(s,a) - \mu^c(s,a) \right| \geq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_n(s,a)}} \right\} \right) \\
&\overset{(a)}{\leq} \sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{t=1}^{n}\sum_{T_n(s,a)=1}^{t} 2\exp\left(-\frac{2T_n(s,a)}{4(\eta^2+\eta)^2} \cdot \frac{4(\eta^2+\eta)^2 \log(SAn(n+1)/\delta)}{2T_n(s,a)} \right) = \delta.
\end{aligned}$$

where, (a) follows from Lemma F.2 when applied for cost. The claim of the lemma follows. □

Corollary 4. *Let the total exploration budget be $n_x = \frac{SA\log(SAn(n+1)/\delta)}{\min_{s,a}\Delta^{c,(2)}(s,a)}$. Define the event $\overline{\xi}_\delta$ as in (58). Then using the exploration policy π_x it can be shown that $\mathbb{P}(\overline{\xi}_\delta) \geq 1-\delta$.*

Proof. Let $n_x = \frac{SA\log(SAn(n+1)/\delta)}{\min_{s,a}\Delta^{c,(2)}(s,a)}$ be the total samples taken for exploration. Let π_e sample each action according to uniform random policy in each state $s \in [S]$. Then the result follows directly from Lemma F.4 in

$$\mathbb{P}\left(\bigcap_{s\in\mathcal{S}} \bigcap_{1\leq a\leq A, T_{n_x}(s,a)\geq 1} \left\{ \left| \frac{1}{T_{n_x}(s,a)} \sum_{t=1}^{T_{n_x}(s,a)} C_t(s,a) - \mu^c(s,a) \right| \geq (2\eta + 4\eta^2)\sqrt{\frac{\log(SAn(n+1)/\delta)}{2T_{n_x}}} \right\} \right) \overset{(a)}{\leq} \delta,$$

where, (a) follows as by noting $T_{n_x} \geq \frac{\log(SAn(n+1)/\delta)}{\min_{s,a}\Delta^{c,(2)}(s,a)}$.

□

G. Additional Experimental Details

In this section we state additional experimental details.

Experiment 1 (Bandit): We implement a bandit environment for $A = 11$ and show that our proposed solution outperforms the safe on-policy and SEPEC (Wan et al., 2022) algorithm. In this experiment we have the $\mu(0) = 0.5, \sigma^2(0) = 10^{-4}$, $\mu(1) = 0.9, \sigma^2(1) = 10^{-4}$ (optimal action), and the sub-optimal actions $a \in \{2, 3, \ldots, 11\}$ have means $\mu(a) \in [0.02, 0.03]$ and high variance $\sigma^2(a) = 40$. Moreover, we set the constraint-value means $\mu^c(a)$ the same as the reward means. The target policy is initialized as $\pi(0) = \pi(1) = 0.4$ while the remaining arms have the 0.2 density evenly distributed among them. So in this environment, the safe on-policy will select the sub-optimal actions less and so reduces MSE at a slower rate. Whereas the SaVeR, complies with the safety constraint and reduces MSE maximally as the number of rounds increases. The performance is shown in Figure 1 (left). Again observe that in Figure 2 (top-left), the oracle keeps the safety budget around 0 and uses all the remaining samples to explore optimally. The SaVeR has a safety budget of almost around 0 as they sample the high cost maximizing action 1 a sufficient number of times to offset the unsafe action pulls. However, safe on-policy and SEPEC again explores the high variance (sub-optimal and unsafe) actions less and has a very high safety budget.

Experiment 2 (Movielens): We conduct this experiment on Movielens dataset for $A = 30$ actions and show that our proposed solution outperforms safe on-policy and SEPEC algorithm. The Movielens dataset from February 2003 consist of 6k users who give 1M ratings to 4k movies. We obtain a rank-4 approximation of the dataset over 128 users and 128

movies such that all users prefer either movies 7, 13, 16, or 20 (4 user groups). The movies are the actions and we choose 30 movies that have been rated by all the users. Hence, this testbed consists of 30 actions and the mean values $\mu(a)$ are the rating of the movies given by the users. and is run over $T = 8000$. The target policy is initialized as $\pi(0) = \pi(1) = 0.4$ while the remaining arms has the 0.2 density evenly distributed among them. We set the cost means $\mu_c(a)$ such that high variance actions have high-cost means. So in this environment, the safe on-policy will select the sub-optimal cost actions less and so reduces MSE at a slower rate as the number of rounds increases. The SEPEC MSE also reduces slower than SaVeR as the number of rounds increases. This is because SEPEC uses an IPW estimator instead of tracking the optimal behavior policy like SaVeR. The SaVeR, complies with the safety constraint and reduces MSE maximally as the number of rounds increases. The performance is shown in Figure 1 (middle-left). Again observe that in Figure 2 (top-right), the oracle keeps the safety budget around 0 and uses all the remaining samples to explore optimally. The SaVeR has a safety budget of almost around 0 as they sample the high reward maximizing action 1 a sufficient number of times to offset the unsafe action pulls. However, safe on-policy and SEPEC again explores the high variance (sub-optimal and unsafe) actions less and has a very high safety budget.

Experiment 3 (Tree): We experiment with a 4-depth 2-action deterministic tree MDP $\mathcal{T}$ consisting of 15 states. In this setting, we have a 4-depth 2-action deterministic tree MDP $\mathcal{T}$ consisting of 15 states. Each state has a low variance arm with $\sigma^2(s, 1) = 0.01$ and high target probability $\pi(1|s) = 0.95$ and a high variance arm with $\sigma^2(s, 1) = 20.0$ and low target probability $\pi(2|s) = 0.05$. Again we set the cost means $\mu^c(a)$ such that high variance actions have high-cost means. Hence, the safe on-policy sampling which samples according to π will sample the second (high variance) arms less and suffer a high MSE. We set $\alpha = 0.25$. We assume that the learner can directly access the $V^{\pi_0}(s_1^1)$ (without any noise) when its safety budget is negative. It can observe $V^{\pi_0}(s_1^1)$ without running any episodic interaction (like Yang et al. (2021). The oracle has access to the model and variances and performs the best. SaVeR lowers MSE comparable to safe onpolicy as the number of episodes increases and eventually matches the oracle's MSE in Figure 1 (middle-right). The SaVeR, oracle, and on-policy have an almost equal safety budget as shown in Figure 2 (bottom-left). Note that we do not run SEPEC in this experiment as it is a bandit algorithm, and the optimization problem of SEPEC do not have a closed form solution in the MDP setting.

Experiment 4 (Gridworld): In this setting we have a 4×4 stochastic gridworld consisting of 16 grid cells. Considering the current episode time-step as part of the state, this MDP is a DAG MDP in which there is multiple paths to a single state. There is a single starting location at the top-left corner and a single terminal state at the bottom-right corner. Let $\mathbf{L}, \mathbf{R}, \mathbf{D}, \mathbf{U}$ denote the left, right, down, and up actions in every state. Then in each state, the right and down actions have low variance arms with $\sigma^2(s, \mathbf{R}) = \sigma^2(s, \mathbf{D}) = 0.01$ and high target policy probability $\pi(\mathbf{R}|s) = \pi(\mathbf{D}|s) = 0.45$. The left and top actions have high variance arms with $\sigma^2(s, \mathbf{L}) = \sigma^2(s, \mathbf{U}) = 0.01$ and low target policy probability $\pi(\mathbf{L}|s) = \pi(\mathbf{U}|s) = 0.05$. We set the cost means $\mu^c(a)$ such that high variance actions have high-cost means. Hence, safe onpolicy which goes right and down with high probability (to reach the terminal state) will sample the low variance arms more and suffer a high MSE. We set $\alpha = 0.25$. Again we assume that the learner can directly access the $V^{\pi_0}(s_1)$ (without any noise) when it's safety budget is negative. It can observe $V^{\pi_0}(s_1)$ without running any episodic interaction (like Yang et al. (2021). SaVeR lowers MSE faster compared to safe onpolicy and actually matches MSE compared to the oracle as well as maintains the safety constraint with increasing number of episodes. We point out that the DAG structure of the Gridworld violates the tree structure under which the oracle and SaVeRbounds were derived. Nevertheless, both methods lower MSE compared to safe onpolicy. Again observe that in Figure 2 (bottom-right), the oracle keeps the safety budget around 0 and uses all the remaining samples to explore optimally. The SaVeRhas a safety budget of almost around 0 as they sample the high reward maximizing action a sufficient number of times to offset the unsafe action pulls. However, safe on-policy again explores the high variance (sub-optimal and unsafe) actions less and has a very high safety budget.

H. Table of Notations

Notations	Definition
s_i^ℓ	State s in level ℓ indexed by i
$\pi(a\|s_i^\ell)$	Target policy probability for action a in s_i^ℓ
$b(a\|s_i^\ell)$	Behavior policy probability for action a in s_i^ℓ
$\sigma^2(s_i^\ell, a)$	Variance of action a in s_i^ℓ
$\widehat{\sigma}_t^{(2),k}(s_i^\ell, a)$	Empirical variance of action a in s_i^ℓ at time t in episode k
$\widehat{\sigma^u}_t^{(2),k}(s_i^\ell, a)$	UCB on variance of action a in s_i^ℓ at time t in episode k
$\mu(s_i^\ell, a)$	Mean of action a in s_i^ℓ
$\widehat{\mu}_t^k(s_i^\ell, a)$	Empirical mean of action a in s_i^ℓ at time t in episode k
$\mu^2(s_i^\ell, a)$	Square of mean of action a in s_i^ℓ
$\widehat{\mu}_t^{(2),k}(s_i^\ell, a)$	Square of empirical mean of action a in s_i^ℓ at time t in episode k
$T_n(s_i^\ell, a)$	Total Samples of action a in s_i^ℓ after n timesteps
$T_n(s_i^\ell)$	Total samples of actions in s_i^ℓ as $\sum_a T_n(s_i^\ell, a)$ after n timesteps (State count)
$T_t^k(s_i^\ell, a)$	Total samples of action a taken till episode k time t in s_i^ℓ
$T_t^k(s_i^\ell, a, s_j^{\ell+1})$	Total samples of action a taken till episode k time t in s_i^ℓ to transition to $s_j^{\ell+1}$
$P(s_j^{\ell+1}\|s_i^\ell, a)$	Transition probability of taking action a in state s_i^ℓ and transition to state $s_j^{\ell+1}$
$M(s_i^\ell) \coloneqq \begin{cases} \\ \\ \end{cases}$	$\sum_a \sqrt{\pi^2(a\|s_i^\ell)\sigma^2(s_i^\ell, a)},$ if $\ell = L$ $\sum_a \sqrt{\sum_{s_j^{\ell+1}} \pi^2(a\|s_i^\ell)\left(\sigma^2(s_i^\ell, a) + P(s_j^{\ell+1}\|s_i^\ell, a)B^2(s_j^{\ell+1})\right)},$ if $\ell \neq L$
$\widehat{M}(s_i^\ell) \coloneqq \begin{cases} \\ \\ \end{cases}$	$\sum_a \sqrt{\pi^2(a\|s_i^\ell)\widehat{\sigma}_t^{(2),k}(s_i^\ell, a)},$ if $\ell = L$ $\sum_a \sqrt{\sum_{s_j^{\ell+1}} \pi^2(a\|s_i^\ell)\left(\widehat{\sigma}_t^{(2),k}(s_i^\ell, a) + P(s_j^{\ell+1}\|s_i^\ell, a)\widehat{B}_t^{(2),k}(s_j^{\ell+1})\right)},$ if $\ell \neq L$

Table 1. Table of Notations